Gender Differences in Early Career Performance Reviews: a Text Mining Study

Shivangi Chopra[1], Lukasz Golab[1]

[1] *University of Waterloo, Canada*

Abstract

It is well known that fewer women than men earn STEM degrees and persist in STEM careers. Since early career experiences affect career attrition, we investigate gender differences in early career performance reviews. Our analysis is enabled by a unique dataset, with nearly 6,000 performance reviews of undergraduate engineering students participating in co-operative internships. Text mining of workplace supervisor comments included in the reviews reveals several gender differences. Male students are more likely to be described as eager, efficient, and independent, whereas female students are perceived as thorough and collaborative. Moreover, male students are more likely to be asked to improve their interpersonal skills, whereas female students are more likely to receive suggestions to improve their business knowledge. Our results thus suggest that men and women are perceived differently in the STEM workplace from the beginning of their careers.

Keywords

gender gap in STEM, co-operative internships, text mining

1. Introduction

The gender gap in Science, Technology, Engineering, and Mathematics (STEM) is well-documented: studies show that fewer women apply to STEM programs [1], obtain STEM degrees [2], and continue with STEM careers [2, 3]. Workplace experiences, especially *early* career experiences, are known to drive career attrition [4, 3]. We therefore ask the following research question: *Are there gender differences in early career performance reviews?*

To answer this question, we analyze workplace performance reviews of students from a large North American university participating in co-operative (co-op) internships. Co-op programs in STEM fields have become popular worldwide, and allow students to alternate between academic study terms and work internships. For many students, co-op internships are the first career experiences in the engineering workplace.

The dataset we analyze consists of nearly 6,000 performance reviews from the 2015/2016 academic year given to undergraduate engineering students. Each review contains two comments: 1) supervisor's feedback on the student's performance, and 2) supervisor's recommendations for the student's future development. Additionally each review includes the student's gender, academic program, and academic level, and we are also given the gender composition of each engineering program at the university. We parse out the words used in these comments and we run statistical tests to identify words with significant frequency differences between the reviews received by male and female students.

We find that male and female students are perceived differently by their co-op employers. Male students are more likely to be described as eager, efficient, and independent, whereas female students are more likely to be described as thorough, dedicated, and collaborative. Besides, male students receive recommendations to improve interpersonal skills and female students are asked to improve their business knowledge. Furthermore, the gender composition of the programs seemed to affect the feedback and recommendations received by the students. The majority gender was more likely to receive technical feedback and recommendations, whereas the minority gender was asked to work on their confidence and ask more questions.

Our results suggest that men and women are perceived differently in the STEM workplace from the beginning of their careers. Whether these gender differences are due to employer perceptions or differences in competencies cannot be determined directly from our data. However, regardless of the underlying reasons, we argue that universities offering co-operative programs should communicate with participating employers to emphasize the importance of unbiased feedback in talent recruitment and retention.

The remainder of this paper is organized as follows. Section 2 summarizes prior work on gender differences in performance reviews. Section 3 describes our dataset and the methodology used to analyze it. Section 4 presents the results. Section 5 summarizes the findings, offers possible explanations for the findings, and presents actionable insights. Finally, Section 6 concludes the paper with directions of future work.

Published in the Workshop Proceedings of the EDBT/ICDT 2022 Joint Conference (March 29-April 1, 2022), Edinburgh, UK

s9chopra@uwaterloo.ca (S. Chopra); lgolab@uwaterloo.ca (L. Golab)

CEUR Workshop Proceedings (CEUR-WS.org)

2. Related Work

We are not aware of any previous work on gender differences in supervisor comments included in early career performance reviews. However, there has been work on gender differences in numeric performance scores given to student interns [5, 6, 7]. The results, however, are inconclusive. A study where technology professionals rated *hypothetical* interns on competence, intelligence, and potential field issues found that men are rated more highly than women [7]. Another study that analyzed evaluations from co-op employers found that female students are rated more highly (than male students) on overall performance as well as on specific criteria including communication, teamwork, and quality of work [5, 6].

More broadly, in the context of postgraduate employment, gender differences in employee (or peer) evaluations have been studied in various fields, including technology, the military, politics, law, sports, and medicine [8, 9, 10, 11, 12, 13, 14, 15, 5, 16]. The evaluations under study were either numeric (ratings), categorical (tags chosen from a predefined list of attributes), or textual. The reported findings are consistent across industries: men receive more actionable and task-oriented feedback and women receive more critical and personality-related feedback.

Among studies that analyzed gender differences in written performance reviews, we found only one that used text mining methods (topic modeling) [8]. This paper studied gender differences in the leadership representation of 146 political leaders by analyzing 1057 comments they received from their colleagues. Other studies that analyzed comments in performance reviews either conducted a qualitative analysis or manually coded the language of the reviews [11, 12, 10, 13, 14, 15, 16]. In those studies, researchers read the comments and coded them according to various parameters, including tone, valence, and skills discussed (technical, communal, agentic, and others). However, one drawback of these studies is the small data size (under 300 performance reviews).

Lastly, we discuss research on gender differences in academic performance reviews. An analysis of 1,224 recommendation letters for postdoctoral fellows in geoscience found that female applicants were only half as likely to receive excellent versus good letters compared to male applicants [17]. These recommendation letters were manually coded in terms of the letter tone and length.

We found no studies in elementary, primary, or secondary education that analyzed gender differences in written performance feedback. However, some studies analyzed gender differences in teacher-student interaction (i.e., verbal feedback) [18, 19, 20, 21, 22]. Studies in STEM classrooms found that teachers tend to attribute boys' success in STEM to ability and boys' failures in STEM to lack of effort, while the opposite is believed to be true for girls [23]. In physical education, male students tend to receive more attention and technical feedback than female students [18, 19]. Studies also found that female students were more likely than male students to internalize the feedback they receive [19]. This internalization of feedback lowered their self-efficacy beliefs and performance [18, 24]. Our study analyzes early career experiences of STEM students to understand if similar differences in feedback persist.

3. Data and Methods

3.1. Data

We analyze three semesters of work performance evaluations, from September 2015 to August 2016, collected by a large North American university. The dataset consists of 5,708 workplace performance reviews of students enrolled in undergraduate engineering co-operative programs. Each review was completed at the end of a four-month internship (in the remainder of this paper, we use the terms 'internship' and 'work term' interchangeably). As part of the evaluation, students receive an overall performance rating that indicates whether the student exceeded, matched, or did not meet the employer's expectations. Hence, we divide students into three categories: above-average, average, and below-average. Along with this overall evaluation rating, the student's supervisor was required to submit short free-text responses to the following questions:

1. *Feedback*: Please comment on the student's overall job performance in terms of their behavioral and developmental performance and expectations with respect to output, quality standards, delivery of goals and assignments.
2. *Recommendations*: Please provide your recommendations for the student's personal and professional development (optional). 42% of the performance reviews have a non-blank recommendation.

Along with this end-of-term performance review, our dataset contains the following information about each student:

1. Gender: male or female,
2. Academic program: one of the 13 engineering programs listed in Table 1, which also shows the gender distribution of each program, sorted by percentage of male students.
3. Seniority: measured in terms of the number of work terms completed: junior students are those who have completed zero or one work terms, and senior students are those who have completed at least four work terms (out of a maximum of six).

The dataset does not include information about the job (for example, job title, company, and location) or the evaluator (for example, position or gender).

We report results for two groups of students: those from programs with less than 40% female students (the first nine in Table 1), and those from programs with greater than or equal to 40% female students (the last four in Table 1)[1]. Table 2 shows the proportions of students in programs with < 40% and $\geq 40\%$ female students and the proportions of students within each group evaluated as below-average, average, and above-average. The table also shows the proportion of male and female students within each group.

3.2. Methods

The goal of this paper is to understand gender differences in written reviews received by student interns. Since these comments have a free-text format, we implemented a parser in Python to convert each comment to a set of standardized word forms (referred to as "words", "tokens", or "terms" in the remainder of the paper). The parser consists of the following standard text mining steps [25]:

1. The text is converted to lower case.
2. Stopwords, which are words that serve a grammatical purpose but do not contain any meaningful information, such as "and", "the" and "is", are removed. Words common in the co-op internship context, including "workterm", "university and "co-op", are also removed.
3. Various forms of certain words and phrases are converted to a common form using regular expression matching[2] (e.g., occurrences of "inter-personal", "interpersonal", and "interpersonal" are converted to "interpersonal", and "hard work" and "hardwork" are converted to "hardwork").
4. Special characters, digits, and punctuation are replaced by white space.
5. Finally, the text is tokenized by white space and stemmed using the NLTK snowball stemmer[3]. Stemming converts words with common meanings but different endings to a common stem. For example, the words "efficient", "efficiently", and "efficiency" are converted to "effici", and "expect", "expected", and "expectation" are converted to "expect".

Table 1
Gender breakdown by program

Program	%Male	%Female
Computer	88%	12%
Mechanical	87%	13%
Mechatronics	86%	14%
Electrical	83%	17%
Software	82%	18%
Nanotechnology	75%	25%
Geological	70%	30%
Civil	67%	33%
System Design	67%	33%
Chemical	60%	40%
Management	58%	42%
Environmental	41%	59%
Biomedical	41%	59%
Total	77%	23%

Then, for each supervisor comment (feedback and recommendations), we conduct a term frequency analysis to identify words that are more frequently used for male students than for female students, and vice versa. We report differences that are statistically significant at a p-value of 0.05 (when using a two-tailed two proportion z-test) and have a statistical power greater than 80%. In addition, for each difference, we report the odds ratio (OR), calculated according to the formula below. The OR indicates the strength (or size) of the difference and can be interpreted as follows. Suppose the odds ratio of token W is 1.5. This means that token W is 1.5 times more likely to occur in Group A (for example, male students) than Group B (for example, female students).

$$Odds\ ratio\ for\ Token\ W\ in\ Group\ A\ versus\ B =$$

$$\frac{\dfrac{\#\ of\ reviews\ in\ Group\ A\ that\ mention\ W}{\#\ of\ reviews\ in\ Group\ A\ that\ do\ not\ mention\ W}}{\dfrac{\#\ of\ reviews\ in\ Group\ B\ that\ mention\ W}{\#\ of\ reviews\ in\ Group\ B\ that\ do\ not\ mention\ W}}$$

We separately report significant gender differences in the feedback and recommendations received by students from programs with < 40% female students and programs with $\geq 40\%$ female students. The analysis is repeated for students with different overall performance ratings (above-average, average and below-average), and seniority levels. To avoid overfitting, we ensure that each group has more than 100 non-blank comments. Common English words with significant differences are excluded from the report for brevity.

[1] We also analyzed the comments received by students from each program separately. However, we observed that the comments received by students in the two groups mentioned above displayed similar trends. Thus, we omit per-program results for brevity.

[2] https://docs.python.org/3/library/re.html

[3] https://www.nltk.org/_modules/nltk/stem/snowball.html

Table 2
Groups based on performance evaluation level

	Programs with < 40% Female students	Programs with $\geq$ 40% Female students
All	86% (82%M, 18%F)	14% (56%M, 44%F)
Above-average	32% (82%M, 18%F)	28% (61%M, 39%F)
Average	47% (80%M, 20%F)	49% (49%M, 51%F)
Below-average	21% (86%M, 14%F)	23% (61%M, 39%F)

4. Results

We now describe the results, treating students in programs with less than 40% female students and those in programs with greater than or equal to 40% female students separately, as mentioned in Section 3.1. Section 4.1 presents word frequency differences in the feedback and recommendations received by male and female students enrolled in programs with < 40% female students. Section 4.2 presents gender differences in word frequencies in feedback and recommendations received by students enrolled in programs with $\geq 40\%$ female students.

4.1. Gender Differences in Programs with < 40% Female Students

4.1.1. Feedback

Table 3 shows the differences in token frequencies in the feedback received by male and female students. On the left, Table 3 shows tokens that are mentioned statistically significantly more frequently in the feedback received by male students. On the right, Table 3 shows the tokens mentioned significantly more frequently in the feedback received by female students.

The lists are sorted by the difference in frequencies, abbreviated Δ, computed as the percentage of male (or female) students whose feedback mentioned a token minus the percentage of female (or male) students whose feedback mentioned this token. For example, feedback received by male students contained “code” 4% more often than feedback received by female students. Asterisks indicate the strength of the statistical significance of the difference, with all reported differences having a p-value of at least 0.05. In addition, Table 3 mentions the odds ratio for each difference. For example, employers are 1.6 times more likely to describe female students as “thorough” than male students.

Even though some of the differences shown in Table 3 appear small in magnitude, they are statistically significant at a p-value of 0.05, have statistical power greater than 80%, and have an odds ratio greater than one (mentioned in Section 3.2).

Feedback received by male students contains more technical terms. Table 3 shows that words relating to technical tasks, including “code”, “tool”, “written”, “hardwar”, “machin”, and “analyz”, are more frequent in the feedback received by male students. Supervisors of male students are four times more likely to refer to them as an “expert”. On the other hand, feedback received by female students mentions their general ability (“profici” in Table 3). This gender difference in the amount of technical feedback received exists in all groups with < 40% female students, irrespective of program, overall evaluation rating, or seniority.

Feedback received by male students contains more mentions of the word “eager”. Manual inspection of the comments containing the token “eager” revealed that these students suggest new ideas and take the initiative to start new tasks. In addition, male students receive feedback on their efficiency and planning (indicated by words such as “effici”, “priori”, “deadlin”, “iter”, and “tackl”).

Table 3 shows that the words “fulltim” and “ecoop” occur more frequently in the feedback received by male students. The token “fulltim” indicates that the employer has extended a full-time job offer to the student. The token “ecoop” refers to a program established by the university under study to allow students to work in their own company (i.e., their start-up) for a co-op work term. Table 3 shows that the token “ecoop” is mentioned in the feedback for 1% of male students and no female students.

Feedback received by female students contains more references to their teamwork and interpersonal skills (indicated by words such as “help”, “collabor”, “delight”, “wonder”, and “joy” in Table 3). In addition, female students receive more feedback on their thoroughness (indicated by words such as “attentiontodetail”, and “thorough” in Table 3), dedication (“dedic”, “enthusiast”), and adaptability (the token “adapt” is mentioned in the feedback received by female students 3.7 times more often than in the feedback received by male students).

Some tokens in Table 3 indicate that male and female students are referred to differently by their employers. Manual inspection of the comments containing the word “addition” indicates that female students are referred to as a “good addition to the team/company”. Manual inspection of the comments containing the word “potenti” indicates that the word is generally used in the context of “has a lot of potential”, and the word “demand” is used to describe a student’s ability to cope with a demanding work environment. These tokens are found more often in the feedback received by female students.

Gender differences in the feedback received by students with different overall performance ratings and seniority levels follow the same trends as above. We omit the details for brevity.

Table 3
Word frequency differences in feedback received by male and female students enrolled in Programs with < 40% female students

Token	Male	Female	Δ	OR	Token	Female	Male	Δ	OR
code	14%	10%	4%***	1.51	help	25%	20%	5%***	1.36
tool	7%	4%	3%**	1.62	dedic	9%	5%	4%***	1.84
fulltim	7%	5%	2%*	1.5	attentiontodetail	7%	4%	3%***	1.94
eager	2%	0%	2%*	2.88	collabor	5%	3%	2%**	1.86
written	3%	1%	2%*	2.55	thorough	6%	4%	2%**	1.63
prioriti	3%	1%	2%**	2.54	enthusiast	5%	3%	2%**	1.79
effici	2%	0%	2%*	6.29	addition	5%	3%	2%**	1.75
hardwar	2%	1%	1%*	2.55	profici	3%	1%	2%***	2.2
machin	2%	1%	1%*	3.07	delight	3%	1%	2%**	2.09
analyz	1%	0%	1%*	4.23	demand	2%	1%	1%**	2.17
expert	1%	0%	1%*	4.16	timemanag	2%	1%	1%***	3.3
deadlin	1%	0%	1%*	6.74	wonder	2%	1%	1%***	2.86
iter	1%	0%	1%*	6.74	adapt	1%	0%	1%***	3.68
ecoop	1%	0%	1%*	inf	joy	1%	0%	1%**	3.01
tackl	3%	2%	1%*	1.85	potenti	1%	0%	1%***	inf

Note. ***: p < .001; **: p < .01; *: p < .05

Table 4
Word frequency differences in recommendations received by male and female students enrolled in Programs with < 40% female students

Token	Male	Female	Δ	OR	Token	Female	Male	Δ	OR
solut	8%	4%	4%**	2.15	allow	8%	3%	5%***	2.88
seek	4%	1%	3%**	3.91	express	4%	0%	4%**	inf
system	4%	1%	3%**	3.21	network	4%	0%	4%***	inf
read	3%	1%	2%*	3.27	oper	5%	1%	4%**	3.13
architectur	3%	1%	2%*	4.75	encourag	7%	5%	4%**	1.55
maintain	3%	1%	2%*	4.41	challeng	9%	5%	4%**	1.89
mistak	2%	0%	2%*	inf	askquestion	9%	5%	4%**	1.93
attent	3%	1%	2%*	3.91	general	4%	1%	3%**	3.53
web	1%	0%	1%*	inf	varieti	3%	0%	3%***	inf
algorithm	1%	0%	1%*	inf	afraid	3%	0%	3%**	3.01
help	1%	0%	1%*	inf	shi	3%	0%	3%***	17.62
cooperat	1%	0%	1%*	inf	explor	4%	1%	3%*	4.05
opinion	1%	0%	1%*	inf	market	3%	1%	2%***	3.34
hear	1%	0%	1%*	inf	tell	1%	0%	1%***	7.2
distract	1%	0%	1%*	inf	comfortzon	1%	0%	1%***	3.62

Note. ***: p < .001; **: p < .01; *: p < .05

4.1.2. Recommendations

Table 4 follows the same format as Table 3 and shows the differences in token frequencies in the recommendations received by male and female students. Again, gender differences in the recommendations received by students with different overall performance ratings and different seniority levels showed similar trends and are not shown for brevity.

Tokens in Table 4 suggest that male students receive more recommendations related to technical skills. This is suggested by words such as "solut" (stem of the word "solution"), "system", "read", "architectur", "maintain", "web", and "algorithm". In addition, male students are recommended to be more attentive to mistakes (indicated by the tokens "attent" and "mistak" in Table 4) and improve their teamwork and interpersonal skills (indicated by "help", "cooperat", "opinion", and "hear").

On the other hand, female students are recommended to "express" themselves, to "network", to not be "afraid" or "shy", and to ask more questions (see Table 4). The recommendations received by female students contains more mentions of the tokens "oper", "general", "varieti",

Table 5
Word frequency differences in feedback received by male and female students enrolled in Programs with $\geq 40\%$ female students

Token	Male	Female	Δ	OR	Token	Female	Male	Δ	OR
abil	22%	14%	8%**	1.71	hardwork	13%	6%	7%**	2.25
understand	20%	12%	8%**	1.81	team	7%	3%	4%**	2.86
littlesupervis	9%	3%	6%***	3.01	applic	6%	2%	4%**	2.89
effici	11%	6%	5%*	2.04	execut	3%	0%	3%*	inf
initi	7%	2%	5%**	3.6	user	3%	0%	3%*	inf
pictur	4%	0%	4%*	inf	technic	3%	0%	3%**	7.09
surpris	5%	1%	4%**	4.35	comprehens	2%	0%	2%**	inf
devic	3%	0%	3%*	inf	writtencomm	2%	0%	2%*	inf
matur	3%	0%	3%*	inf	expertis	2%	0%	2%*	8.93
prioriti	3%	0%	3%*	inf	smart	1%	0%	1%*	inf
newtask	3%	0%	3%**	10.67	stack	1%	0%	1%*	inf
growth	2%	0%	2%*	inf	legaci	1%	0%	1%*	inf
difficulti	1%	0%	1%*	inf	style	1%	0%	1%*	inf
persist	1%	0%	1%*	inf	joy	1%	0%	1%*	inf
ecoop	1%	0%	1%*	inf	read	1%	0%	1%*	inf

Note. ***: p < .001; **: p < .01; *: p < .05

"explor", and "market" (see Table 4). Manual inspection of comments containing these tokens reveals that female students receive more recommendations to explore and increase their variety of knowledge, especially about business operations.

Table 4 indicates that recommendations received by female students contained more occurrences of the words "allow", "encourag", "challeng", and "comfortzon". Manual inspection of comments containing these tokens suggests that female students were encouraged to challenge themselves and leave their comfort zones more often than male students.

4.2. Gender Differences in Programs with $\geq 40\%$ Female Students

Tables 5 and 6 list the differences in word frequencies in the feedback and recommendations, respectively, received by students enrolled in programs with $\geq 40\%$ female students. These tables follow the same format as Tables 3 and 4. Again, we omit gender differences in groups based on overall performance ratings and seniority, which show similar trends.

4.2.1. Feedback

Table 5 indicates that comments received by female students are more related to technical performance (suggested by tokens such as "applic", "execut", "user", "technic", "writtencomm", "stack", and "read"). In addition, tokens such as "expertis" and "legaci" are found more frequently in the feedback received by female students. On the other hand, feedback received by male students references their "ability". This is in contrast to the results presented in Section 4.1, where male students received more technical feedback than female students.

Nevertheless, some of the feedback received by male students is similar to the feedback received by male students from programs with < 40% female students (Section 4.1). Male students are more likely to receive feedback on their eagerness to start new tasks (suggested by the tokens "newtask" and "initi" in Table 5, where "initi" is the word stem for "initiate" and "initiative"). They are also more likely to receive feedback on their planning and efficiency ("effic", "pictur", "prioriti"). The token "littlesupervis" in Table 5 indicates that supervisors find male students to be more independent than female students.

Table 5 indicates that female students received more feedback on their hard work, thoroughness ("comprehens", which is the word stem for "comprehensive"), teamwork, and interpersonal skills. Female students from programs with < 40% female students received similar feedback from their employers (see Section 4.1).

Feedback given to male students contains more mentions of the words "surpris", "growth", "persist", "difficulti", and "matur" (see Table 5). Manual inspection of comments containing these terms revealed that these employers were pleasantly surprised to see the students' growth, persistence, and maturity.

Finally, similar to programs with < 40% female students (Section 4.1), the token "ecoop" is mentioned for 1% of male students and no female students.

Table 6
Word frequency differences in recommendations received by male and female students enrolled in Programs with $\geq 40\%$ female students

Token	Male	Female	Δ	OR	Token	Female	Male	Δ	OR
say	4%	0%	4%*	inf	oper	5%	1%	4%*	8.81
mistak	3%	0%	3%*	inf	creativ	5%	1%	4%*	8.81
reserv	3%	0%	3%*	inf	surround	4%	0%	4%*	inf
team	3%	0%	3%*	inf	knowledg	4%	0%	4%*	inf
public	3%	0%	3%*	inf	instinct	3%	0%	3%*	inf
speak	3%	0%	3%*	inf	quick	3%	0%	3%*	inf
open	3%	0%	3%*	inf	generat	3%	0%	3%*	inf
expect	2%	0%	2%*	inf	difficult	3%	0%	3%*	inf
distract	2%	0%	2%*	inf	system	3%	0%	3%*	inf
error	2%	0%	2%*	inf	learn	3%	1%	2%***	19.33
topic	2%	0%	2%*	inf	document	2%	0%	2%*	inf
softskil	2%	0%	2%*	inf	explor	2%	0%	2%*	inf
listen	2%	0%	2%*	inf	interest	1%	0%	1%*	inf
respect	2%	0%	2%*	inf	compani	1%	0%	1%**	2.02
complex	2%	0%	2%**	inf	deal	1%	0%	1%***	4.97

Note. ***: p < .001; **: p < .01; *: p < .05

4.2.2. Recommendations

Table 6 indicates that male students are referred to as "reserved" and are recommended to "speak" (suggested by tokens such as "reserv", "say", "public", "speak", and "open"). This is in contrast to the results reported in Section 4.1, where female students were recommended to ask more questions.

Table 6 also indicates that female students receive more technical recommendations than male students. Tokens such as "creativ", "knowledg", "generate", "system", "interest", "document", and "learn", are more common in the recommendations received by female students. On the other hand, recommendations received by male students contain more occurrences of the tokens "topic" and "complex". Again, this is in contrast to the results shown in Section 4.1, where male students received more technical recommendations.

Nevertheless, some recommendations given to students in programs with $\geq 40\%$ female students are similar to those given to students in programs with < 40% female students (Section 4.1). For example, similar to male students from programs with < 40% female students (Section 4.1), male students from programs with $\geq 40\%$ female students are also recommended to keep an eye out for mistakes (indicated by "mistak", "distract", "error" in Table 6) and improve their teamwork and interpersonal skills ("team", "softskill", "listen", "respect"). Female students from programs with $\geq 40\%$ female students are recommended to gain operational knowledge (indicated by "oper", "surround", "explor", and "compani" in Table 6 and confirmed by manual inspection of the comments containing these tokens). The same recommendations were received by female students from programs with < 40% female students.

5. Discussion

The main findings of this study and their significance are as follows.

Observation #1: We found the following gender differences in all groups of students, irrespective of the overall performance rating, seniority, and the gender composition of their academic programs.

1. Female students are more likely than male students to be appreciated for their thoroughness, dedication, enthusiasm, hard work, adaptability, teamwork, and interpersonal skills.
2. Male students are more likely than female students to be appreciated for their eagerness, planning, efficiency, and independence.
3. Female students are recommended to increase their business knowledge, including general information about the market and company operations.
4. Male students are recommended to keep an eye out for mistakes and improve their teamwork and interpersonal skills.

These gender differences in feedback and recommendations may be due to gender differences in (a) how employers perceive their students' competencies, (b) opportunity, or (c) students' abilities.

Gender differences in perceived competencies: The gender differences we found are consistent with past studies that examined feedback in education and in the workplace. For example, studies examining professionals in technology, military, politics, and law found that women were appreciated for their communal qualities (e.g., those related to social relationships) and men were appreciated for their agentic qualities (e.g., those related to goal achievement) [8, 9, 10, 11, 12, 13, 14]. In addition, women were more often tagged as "enthusiastic", "organized", and "unaware" and men as "analytical", "dependable", and "irresponsible" [9]. Studies in STEM classrooms indicate that teachers attribute male students' achievements to their ability, and female students' achievements to their hard work [23]. Social scientists and psychologists confirm the existence of stereotypes of men and women [26, 27]. Therefore, a possible reason behind the gender differences we found may be the unconscious gender bias of the evaluator (i.e., the work term supervisor).

Studies suggest that positive and negative gender stereotypes found in evaluations affect students' self-image and career choices [26, 28, 29, 24, 19]. Additionally, experiments found that gendered language in performance evaluations may affect hiring and promotion decisions [14, 9]. For example, when conducting a blind review of candidates for promotion, participants chose candidates described as "good at taking initiative". Since these (agentic) characteristics occur in the performance evaluations of men more often than women, this may lead to fewer promotion opportunities for women. Additionally, participants considered collaborative skills, and thus, female profiles, less suitable for leadership roles [14]. Overall, since task-oriented qualities are more valuable to an organization than social-oriented qualities [30], the gender stereotypes in performance evaluations may give men a better chance to be hired, promoted, and more highly paid.

More female than male students leave STEM programs and careers [31, 2]. Potential reasons for this include sexism in teams, the masculine culture in the STEM education and workplace, and dissatisfaction over pay and promotion opportunities [3]. Therefore, eliminating gender bias from early career performance reviews can help plug the "leaky" pipeline. In particular, universities offering co-op programs should communicate with participating co-op employers to emphasize the importance of unbiased feedback. One problem with implicit bias is that many people are not aware that they are biased, emphasizing the importance of diversity training for workplace supervisors.

Gender differences in opportunity: We found that female students were appreciated for their adaptability more often than male students, indicating that perhaps female students were initially perceived to be more incompatible with the company culture. Past studies suggest that the masculine work and after-work culture of male-dominated professions make women uncomfortable [32]. This masculine culture may cause female students to consciously or unconsciously limit their workplace interactions (with peers and supervisors), limiting their access to operational knowledge. Given fewer female supervisors [2], female students may have found it difficult to communicate within a male-dominated hierarchy.

Gender differences in ability: Biological or society-driven differences in ability may have led to the gender differences in performance evaluations reported in this study. Past studies found that females were more likely to possess *both* high mathematical and verbal abilities and males were more likely to demonstrate higher mathematical abilities relative to their verbal abilities [28]. In addition, studies found that female students preferred people-oriented roles [33], displayed more altruistic tendencies [1], scored higher on teamwork and interpersonal communication [5, 6], and outperformed male students at collaborative problem solving tasks [34].

Observation #2: There appears to be a relationship between the gender composition of academic programs and the comments received by students in those programs. This is particularly noteworthy because it occurs in a field with (traditionally) pro-male ability beliefs. We found that in programs with < 40% female students, a higher proportion of male students received feedback on their technical performance in comparison to female students. The recommendations received by male students also contained more technical directions for improvement. On the other hand, female students were recommended to participate, be less shy, and ask more questions. For programs with $\geq 40\%$ female students, the opposite is true. In these programs, female students receive more technical feedback and recommendations, and male students are recommended to be less reserved and speak more openly. This trend exists across all groups of students, irrespective of overall performance scores and seniority.

Gender differences in technical evaluation: The above observation is consistent with past observational studies that analyzed gender differences in teacher-student interaction and the feedback received by secondary school students. Some studies found that male students received more attention and feedback, particularly praise, criticism, and technical information, irrespective of the subject being taught (sports, modern languages, mathematics, science, and humanities) [19, 18, 20]. However, this was reversed in classes that contained as many or more female students [20]. Since feedback and recommendations on technical and behavioral skills are important for co-op students [30], universities may want to ensure that co-op evaluation forms include explicit requests to comment on students' technical skills.

Studies that analyzed the performance reviews of men and women in (a) technology and professional-services firms [14], (b) a leadership development program [8], and (c) navy academy students [9], found more mentions of technical words in the feedback received by men than women. These gender differences in technical feedback were attributed to the pro-male ability bias that exists in these fields. However, since all of these studies investigated samples containing less than 25% women, our results suggest the need for further investigation.

Gender differences in participation: A study conducted in a secondary school reported that both male and female students participated more when their own gender was the majority gender in the classroom [20]. This was found irrespective of the subject being taught. Similarly, a study where engineering students were randomly assigned to teams (or "micro-environments") with varying gender composition reported similar conclusions. This study found that when female students were the minority in a team (less than 25%), they spoke less, were less involved in teamwork, and felt less confident than female students assigned to teams where they were in the majority (75% or more) [35]. This was true regardless of the students' academic seniority. Moreover, female students from male-majority teams reported lowered engineering career aspirations after the team interaction [35].

Past studies attribute the reason behind this difference in participation to isolation (or social-belongingness concerns) and stereotype threat (the concern that one will be judged in terms of a stereotype) [20, 35]. Female students were more affected by the gender composition in a classroom, leading to recommendations to create single-sex or gender-parity micro-environments (e.g., in-class teams or study groups) [35, 20]. Researchers experimenting with varying proportions of male and female students in engineering teams found that gender-balanced micro-environments are particularly important for first-year students, to ensure these students do not lose confidence and drop out of STEM fields [35]. Gender-balanced micro-environments helped students focus on learning, participate more freely, and in turn, gain the confidence to persist in gender-imbalanced environments. Another study found that participation in social-belonging interventions during student orientation programs improved female students' social attitude and academic performance in male-dominated STEM programs [36].

Our results similarly suggest that co-op students working in environments where they are not the majority gender participate less in team activities and may need additional encouragement. As suggested by past studies, gender imbalanced classrooms and workplaces may experiment with social-belonging interventions and gender-parity micro-environments and note their effect on student confidence.

Observation #3: Different words were used to describe the minority and the majority gender. Phrases including "has a lot of potential", "challenge yourself", "allow yourself to grow", and "come out of your comfort zone", are more common in the comments received by female students from programs with < 40% female students. On the other hand, phrases including "surprised by performance" and "mature" are more common in the comments received by male students from programs with $\geq 40\%$ female students.

Studies of tokenism support the above observation and suggest that bias against a group occurs when said group is a minority in *any* given field [37]. Related work on minority groups (in terms of race and gender) presents conflicting reports on whether the feedback provided to those groups is more lenient or harsh [11, 12, 38]. However, most studies that report gender differences in feedback note that the same trait is described more positively for men than for women [8, 11, 12, 13, 15]. Note that all these studies were conducted in male-dominated professions.

6. Conclusions

In this paper, we analyzed gender differences in early career workplace performance reviews. To do so, we used a unique dataset corresponding to work term evaluations of students enrolled in engineering co-operative programs. We used text mining methods to analyze word frequency differences in employer feedback and recommendations for professional development.

We found that male students were appreciated for taking initiative more often than female students. They were described as efficient and independent and were recommended to improve their interpersonal and teamwork skills. On the other hand, female students were appreciated for being thorough, hardworking, social, and collaborative. They were advised to gain business knowledge more often than male students. We also found differences in the comments received by students in male versus female-dominated programs. We found that in both groups of engineering programs, the majority gender received more technical feedback and recommendations, and the minority gender was advised to ask more questions and be more confident.

Our main takeaway message is that men and women appear to be perceived differently in the STEM workplace from the beginning of their careers. Since reiteration of gendered feedback leads to career dissatisfaction and attrition [24, 14, 3], our results emphasize the importance of unbiased feedback in early career settings such as co-operative internships. Moreover, since our results suggest a possible link between the gender composition of the programs and the feedback received by the majority and minority gender, special attention should be paid to encourage minority groups.

The results presented in this paper should be interpreted carefully since they are based on data from a single North American institution. Nevertheless, we believe that our data-driven study is a useful starting point for further analysis. For example, an interesting direction for future work is to interview STEM alumni to determine if their co-op experiences affected their career paths. Furthermore, it may be useful to investigate the effect of the workplace supervisor's gender on performance reviews (we were unable to do this analysis because our dataset did not include any information about workplace supervisors).

References

[1] S. Chopra, H. Gautreau, A. Khan, M. Mirsafian, L. Golab, Gender differences in undergraduate engineering applicants: A text mining approach, in: Proceedings of the 11th International Conference on Educational Data Mining, EDM 2018, Buffalo, NY, USA, July 15-18, 2018, 2018, pp. 44–54.

[2] A. Perreault, Analysis of the distribution of gender in stem fields in canada, http://wiseatlantic.ca/wp-content/uploads/2018/03/WISEReport2017_final.pdf, ???? Accessed: 20th March, 2019.

[3] J. Hunt, Why do women leave science and engineering?, ILR Review 69 (2016) 199–226.

[4] A. Kauhanen, S. Napari, Gender differences in careers, Annals of Economics and Statistics (2015) 61–88.

[5] S. Chopra, A. Khan, M. Mirsafian, L. Golab, Gender differences in work-integrated learning assessments., in: Proceedings of the International Conference on Educational Data Mining (EDM), 2019, pp. 524–527.

[6] S. Chopra, A. Khan, M. Mirsafian, L. Golab, Gender differences in work-integrated learning experiences of stem students: From applications to evaluations, International Journal of Work-Integrated Learning 21 (2020) 253–274.

[7] E. D. Reilly, K. R. Rackley, G. H. Awad, Perceptions of male and female stem aptitude: The moderating effect of benevolent and hostile sexism, Journal of Career Development 44 (2017) 159–173.

[8] E. Doldor, M. Wyatt, J. Silvester, Statesmen or cheerleaders? using topic modeling to examine gendered messages in narrative developmental feedback for leaders, The Leadership Quarterly 30 (2019) 101308.

[9] D. G. Smith, J. E. Rosenstein, M. C. Nikolov, D. A. Chaney, The power of language: Gender, status, and agency in performance evaluations., Sex Roles 80 (2019).

[10] L. H. Keith, Visibility invisibility: Feedback bias in the legal profession, J. Gender Race & Just. 23 (2020) 315.

[11] P. Cecchi-Dimeglio, How gender bias corrupts performance reviews, and what to do about it, Harvard Business Review 12 (2017).

[12] K. Snyder, The abrasiveness trap: High-achieving men and women are described differently in reviews, Fortune Magazine 26 (2014) 08–14.

[13] S. J. Correll, K. R. Weisshaar, A. T. Wynn, J. D. Wehner, Inside the black box of organizational life: The gendered language of performance assessment, American Sociological Review 85 (2020) 1022–1050.

[14] R. Silverman, Gender bias at work turns up in feedback, 2015. URL: https://www.wsj.com/articles/gender-bias-at-work-turns-up-in-feedback-1443600759.

[15] K. Brucker, N. Whitaker, Z. S. Morgan, K. Pettit, E. Thinnes, A. M. Banta, M. M. Palmer, Exploring gender bias in nursing evaluations of emergency medicine residents, Academic Emergency Medicine 26 (2019) 1266–1272.

[16] A. S. Mueller, T. M. Jenkins, M. Osborne, A. Dayal, D. M. O'Connor, V. M. Arora, Gender differences in attending physicians' feedback to residents: a qualitative analysis, Journal of Graduate Medical Education 9 (2017) 577–585.

[17] K. Dutt, D. L. Pfaff, A. F. Bernstein, J. S. Dillard, C. J. Block, Gender differences in recommendation letters for postdoctoral fellowships in geoscience, Nature Geoscience 9 (2016) 805.

[18] V. Nicaise, G. Cogérino, J. Bois, A. J. Amorose, Students' perceptions of teacher feedback and physical competence in physical education classes: Gender effects, Journal of teaching in Physical Education 25 (2006) 36–57.

[19] V. Nicaise, J. E. Bois, S. J. Fairclough, A. J. Amorose, G. Cogérino, Girls' and boys' perceptions of physical education teachers' feedback: Effects on performance and psychological responses, Journal of sports sciences 25 (2007) 915–926.

[20] S. Drudy, M. Ú. Chatháin, Gender effects in classroom interaction: Data collection, self-analysis and reflection, Evaluation & Research in Education 16 (2002) 34–50.

[21] P. C. Burnett, Teacher praise and feedback and students' perceptions of the classroom environment, Educational psychology 22 (2002) 5–16.

[22] M. G. Jones, J. Wheatley, Gender differences in teacher-student interactions in science classrooms, Journal of research in Science Teaching 27 (1990) 861–874.

[23] J. Tiedemann, Gender-related beliefs of teachers in elementary school mathematics, Educational studies in Mathematics 41 (2000) 191–207.

[24] M. Mayo, M. Kakarika, J. C. Pastor, S. Brutus, Aligning or inflating your leadership self-image? a longitudinal study of responses to peer feedback in mba teams, Academy of Management Learning &

Education 11 (2012) 631–652.

[25] W. B. Croft, D. Metzler, T. Strohman, Search engines: Information retrieval in practice, volume 520, Addison-Wesley Reading, 2010.

[26] M. E. Heilman, Gender stereotypes and workplace bias, Research in organizational Behavior 32 (2012) 113–135.

[27] J. Lorber, S. A. Farrell, et al., The social construction of gender, Sage Newbury Park, CA, 1991.

[28] M.-T. Wang, J. L. Degol, Gender gap in science, technology, engineering, and mathematics (stem): Current knowledge, implications for practice, policy, and future directions, Educational Psychology Review 29 (2017) 119–140. URL: https://doi.org/10.1007/s10648-015-9355-x. doi:10.1007/s10648-015-9355-x.

[29] N. Dasgupta, J. G. Stout, Girls and women in science, technology, engineering, and mathematics: Steming the tide and broadening participation in stem careers, Policy Insights from the Behavioral and Brain Sciences 1 (2014) 21–29.

[30] R. K. Coll, K. E. Zegwaard, Perceptions of desirable graduate competencies for science and technology new graduates, Research in Science & Technological Education 24 (2006) 29–58.

[31] D. Hango, Gender differences in science, technology, engineering, mathematics, and computer science (STEM) programs at university, Insights on Canadian Society (2013).

[32] C. Seron, S. S. Silbey, E. Cech, B. Rubineau, Persistence is cultural: Professional socialization and the reproduction of sex segregation, Work and Occupations 43 (2016) 178–214.

[33] R. Su, J. Rounds, P. I. Armstrong, Men and things, women and people: a meta-analysis of sex differences in interests., Psychological bulletin 135 (2009) 859.

[34] OECD, Collaborative problem solving (2017). URL: https://www.oecd-ilibrary.org/content/paper/cdae6d2e-en. doi:https://doi.org/https://doi.org/10.1787/cdae6d2e-en.

[35] N. Dasgupta, M. M. Scircle, M. Hunsinger, Female peers in small work groups enhance women's motivation, verbal participation, and career aspirations in engineering, Proceedings of the National Academy of Sciences 112 (2015) 4988–4993.

[36] G. M. Walton, C. Logel, J. M. Peach, S. J. Spencer, M. P. Zanna, Two brief interventions to mitigate a "chilly climate" transform women's experience, relationships, and achievement in engineering., Journal of Educational Psychology 107 (2015) 468.

[37] R. M. Kanter, Some effects of proportions on group life, in: The gender gap in psychotherapy, Springer, 1977, pp. 53–78.

[38] K. D. Harber, Feedback to minorities: Evidence of a positive bias., Journal of personality and social psychology 74 (1998) 622.