What Women Want: Analyzing Research Publications to Understand Gender Preferences in Computer Science

Rada Mihalcea, Charles Welch
Computer Science and Engineering
University of Michigan
{mihalcea,cfwelch}@umich.edu

Abstract

While the number of women who choose to pursue computer science and engineering careers is growing, men continue to largely outnumber them. In this paper, we describe a data mining approach that relies on a large collection of scientific articles to identify differences in gender interests in this field. Our hope is that through a better understanding of the differences between male and female preferences, we can enable more effective outreach and retention, and consequently contribute to the growth of the number of women who choose to pursue careers in this field.

1 Introduction

The gender gap in computer science and engineering (CSE) has significant implications on progress made in this field: computer scientists are responsible for many of the aspects of our daily life – with important applications ranging from search engines and large-scale social media sites, to educational applications and health-care technologies to managing economic systems, trading stocks, and national security – but women are often not involved in the development of these products, which can sometime make them less appropriate for the women consumers.

Recent statistics indicate only 19% of engineering Bachelor degrees were awarded to women in 2012 and fewer than 20% of enrolled engineering undergraduates were women (American Society of Engineering Education 2013), and these figures are alarmingly low. National Science Foundation statistics (National Science Foundation 2012) show that 28% of computer science bachelor degrees awarded in 2000 were given to women and this has decreased to 18% in 2009.

In this paper, we examine the interests of men and women in the field of CSE. By analysing a large number of publications in this field, we identify the areas that are preferred by each gender, and also how these preferences appear to change over time. We believe such findings are important not only for the purpose of having a better understanding of our interests and preferences in this field, but also as a critical piece of information that can be used for more effective recruitment and retention of women in CSE.

Specifically, the paper makes three important contributions. First, we compile what we believe is the first and largest collection of CSE publications annotated with author gender. In the process, we also explore techniques for semi-automatic gender annotation, which we use to label the gender of the authors in our collection.

Second, we identify the areas that are primarily preferred by men and women. Starting with the collection of CSE publications annotated for gender, we find a set of papers that are assigned to one or more categories using the Association for Computing Machinery (ACM) computing classification system. We provide data and analysis of the number of men and women who published papers in a set of eleven categories and a large set of sub-categories.

Third, we perform an initial temporal analysis of the trends of these preferences over time, which can provide insight into how interests have shifted with the growth or change associated with specific areas within CSE.

2 Related Work

The gender gap in CSE has been consistently observed over the past two decades (Spertus 1991; Margolis and Fisher 2003; Beauboef and Zhang 2011). While there is no definite explanation of the gender differences noticed in CSE in particular and in science and engineering in general, research in education, psychology, and sociology has identified a number of factors. For instance, a recent study in social psychological research (Cech et al. 2011) proposed two possible explanations as to why women are underrepresented in science and engineering. First, women might not consider science and engineering areas because they believe that the demands of such careers may interfere with their family plans. Second, women are also believed to leave due to low self-assessment of their skills in tasks required in science and engineering careers.

In a study concerned with the use of real-world contexts in computer science courses, (Burn and Holloway 2006) describe an experiment involving two sections of an introduction to a CSE course, one using a "traditional" approach and one emphasizing real-world contexts. Students perceptions of the importance and relevance of the course material were markedly greater in the "real-world" course. Moreover, gaps in academic performance on exams and homework assignments between males and females and between students of color and white students were smaller in the section integrating real-world contexts.

A number of reports have also indicated the need to improve the quality of science, technology, engineering, and mathematics education to support a diverse student body and prepare engineers to be competitive in a global work force (National Academy of Engineering 2004; National Academy of Sciences, National Academy of Engineering, Institute of Medicine 2007; National Science Board 2004). Research such as the pivotal work of (Seymour and Hewitt 1997) and (Tobias 1990) has demonstrated that, in many cases, faculty teaching practices can greatly affect the quality of education in these fields. Specifically, such practices can have a direct impact on student achievement (e.g., student involvement, engagement, knowledge construction, and cognitive development) and, as a result, on student decisions to persist in engineering (Angelo and Cross 1993; Blackburn and Lawrence 1995). (Tobias 1990) and others (Claxton and Murrell 1987; Felder 1993) note that introductory science courses are often responsible for driving off many students who have an initial intention and the ability to earn science degrees but instead switch to nonscientific fields (i.e., students in the second tier). Women and other underrepresented minorities are over-represented in this population. (Felder 1993) summarizes several teaching practices that contribute to the departure of second tier students from engineering. These include a lack of classroom community, a lack of identifiable goals in a course, relegation of students to almost complete passivity in the classroom, and failure to motivate interest in science by establishing its relevance to the students' lives and personal interests.

It is believed that there are a set of myths and misconceptions that affect how students perceive computer science (Beauboef and McDowell 2008). These misconceptions including low social skills or interaction with others, lower retention rates for certain groups of students, when in fact almost all industry developed code is written collaboratively. A study on pair-programming with women in introductory computer science courses at the University of California Santa Cruz states that pair-programming increases retention rates (Werner and McDowell 2013). An effort at Swarthmore college to increase enrollment in introductory computer science courses has been accomplished with a change in curriculum combined with a student mentoring program; this program has made the student body more cooperative and connected, with an increase in student groups such as *Women in CS* (Newhall et al. 2014).

Unlike these previous case studies, we believe our paper provides the first large-scale principled exploration of women interests in CSE areas, thus leading to a better understanding of what is behind the women's motivation to pursue a career in CSE, and also enabling the portability of these studies to other fields in engineering.

3 Building a Collection of Papers Labeled for Gender

In order to infer gender interests for the field of CSE, we need a large collection of publications that cover all the computer science areas, and which have their authors annotated for gender. We build our collection using the ACM digital library, and consequently use the ACM computing categories.

3.1 ACM Computing Classification System

Our system of categories is based on the ACM's Digital Library Computing Classification System (CCS), which contains various levels organized in a tree structure, with the specificity of a category in the tree being proportional to node depth. We are using parts of the first two levels of categories in this tree, including a first level of eleven broad categories, which are in turn split into 79 subcategories. Table 1 shows the categories and sub-categories that we use in our work, and for which we collect papers from the ACM Digital Library.

3.2 Crawling ACM Publications

A web crawler was written to gather publications for the 11 categories and 79 sub-categories from the CCS tree. We collected information on a total of 4,484 papers. For each paper, the web crawler extracts several pieces of information, including the title, the authors, the institution of the authors, the abstract, the CCS categories for the paper. Although we aimed to collect 100 papers for each category, there are categories for which the crawler could not find 100 publications. Note also that one paper typically belongs to multiple categories. On average, each category includes 1131 papers, and each sub-category includes 211 papers.

3.3 Finding the Gender of a Name

A critical piece of information required by our work is the gender of the authors of the research papers in the collection. Thus, an important challenge that we faced was the identification of the gender of the author names. While this may be a relatively trivial task for names for which there is a good Census data (e.g., American names), the problem becomes significantly more challenging when the names spread a large number of cultures as is the case with the authors of the ACM publications.

The entire dataset consists of 9,644 authors, with 4,388 unique first names. Given the large number of names in our collection, full manual annotation is not an option, and therefore we use a combination of techniques along with crowdsourced annotation to handle those cases that cannot be covered by automatic methods.

To determine the accuracy of the individual methods, and also to decide on the best way to combine these methods, we built a list of 100 "difficult" names. The list is compiled using uncommon names that could not be found in the knowledge sources we consider or which cannot be labeled by the methods we use (see Table 2 for a listing of the knowledge sources and methods we use). For instance, the following ten names are included on our "difficult" list: *Jayanth, Jifeng, Dalin, Aniket, Kossi, Rivalino, Tadashi, Venkatesh, Xiaoyu, Zibin*. This list of names is manually annotated by looking up the authors personal website or information provided by the university or institution they worked with to determine the gender.

Category	Sub-categories
Hardware (668)	Printed circuit boards, Communication hardware, interfaces and storage, Integrated circuits, Very large scale integration design, Power and energy, Electronic design automation, Hardware validation, Hardware test, Robustness, Emerging technologies
Computer Systems Organization (880)	Architectures, Embedded and cyber-physical systems, Real-time systems, Dependable and fault-tolerant systems and networks
Networks (946)	Network architectures, Network protocols, Network components, Network algorithms, Network performance evaluation, Network properties, Network services, Network types
Software (1234)	Software organization and properties, Software notations and tools, Software creation and management
Theory (1550)	Models of computation, Formal languages and automata theory, Computational complexity and cryptography, Logic, Design and analysis of algorithms, Randomness geometry and discrete structures, Theory and algorithms for application domains, Semantics and reasoning
Mathematics of Computing (1261)	Discrete mathematics, Probability and statistics, Mathematical software, Information theory, Mathematical analysis, Continuous mathematics
Information Systems (1379)	Data management systems, Information storage systems, Information systems applications, World Wide Web, Information retrieval
Security and Privacy (767)	Cryptography, Formal methods and theory of security, Security services, Intrusion/analomy detection and malware mitigation, Security in hardware, Systems security, Network security, Database and storage security, Software and application security, Human and societal aspects of security and privacy
Human-Centered Computing (792)	Human computer interaction (HCI), Interaction design, Collaborative and social computing, Ubiquitous and mobil computing, Visualization, Accessibility
Computing Methodologies (1654)	Symbolic and algebraic manipulation, Parallel computing methodologies, Artificial intelligence, Machine learning, Modeling and simulation, Computer graphics, Distributed computing methodologies, Concurrent computing methodologies
Applied Computing (1309)	Electronic commerce, Enterprise computing, Physical sciences and engineering, Life and medical sciences, Law social and behavioral sciences, Computer forensics, Arts and humanities, Computers in other domains, Operations research, Education, Document management and text processing

Table 1: Categories and sub-categories in the ACM Computing Classification System

We identified five techniques to automatically annotate name gender. The first two are simple heuristics, based on the presence of the name in a large collection of names. We use the U.S. Census Data,[1] and the dataset from Tang et al (Tang et al. 2011). Additionally, we also use three methods that are available as Web services: the Wolfram Alpha engine, GPeters.com,[2] and GenderGuesser.com. Table 2 shows the accuracy of each of these annotation methods on the dataset of 100 difficult names. Note that the accuracy is calculated on the entire set of 100 names, regardless of actual coverage, and therefore methods are penalized for names that they do not cover.

We also combine the gender annotation techniques using three meta-classifiers. The first meta-classifier uses majority voting, and chooses the most common gender suggested by the five techniques we use. Second, we use a pipeline where the cases not covered by GenderGuesser are annotated using the Tang names heuristic, followed by an assumption of male gender for all the names that are left unannotated. Finally, we use a pipeline similar to the one before, but replacing the male default annotation with a manual annotation through crowdsourcing. The results of these meta-classifiers are also included in Table 2.

Method	Accuracy
Wolfram Alpha	6%
GPeters.com	49%
GenderGuesser.com	75%
US Census Data	3%
Tang Names	32%
Majority	81%
GenderGuesser/Tang/Male	84%
GenderGuesser/Tang/Manual	87%

Table 2: Accuracy for gender labeling on 100 difficult names

What we learn from these evaluations is that the highest accuracy can be obtained by using a pipeline consisting of GenderGuesser followed by a heuristic based on the Tang dataset, and finally followed by crowdsourcing to address the names that are not covered by these two methods.

Applying this pipeline on the entire set of 4,388 unique names results in 381 names that cannot be annotated neither by GenderGuesser nor by the Tang-based heuristic. We use Amazon Mechanical Turk (AMT) to annotate the gender of these 381 names. The names are separated into twenty tasks (or hits, in AMT terminology). Each task contains eighteen or nineteen names whose gender is unknown, and also one name which was manually annotated for gender by us and which we use as a check against spamming. For each name, we asked the following question: *Given the following first (given) name, determine if the name is more often given to males or females.* We requested that the names be annotated by workers with a 92% approval rate or higher who had at least a few hundred accepted hits. If the known name in a hit was wrongly annotated by a worker, the entire annotation from that worker was discarded, and the hit was reposted for annotation by another worker. Each name was annotated by three different workers, and the most frequent gender is used as a final label.

Name	Annotated Gender
Sergio, Alessandro, Anurag, Chun-Chuan, Dimitris, Milica, Emilio, Anish, Zoltán, Vasileios	Male
Elizabeth, Archana, Eleanor, Rui, Pearl, Judith, Liliana, Ariel, Nirattaya, Soultana	Female

Table 3: Randomly sampled names labeled for gender

Table 3 shows ten randomly sampled male and female names from our final dataset. As a posthoc evaluation of the accuracy of the gender assignment, we randomly select 100 names from the set of 4,388 first names and manually ver-

[1] http://www.census.gov/data.html

[2] http://www.gepeters.com/names/baby-names.php

ify the correctness of the label. This evaluation resulted in 92% accuracy, which we believe is acceptable for use in the gender analyses that we want to perform on this dataset.

4 Gender in Computer Science

The final dataset used for our gender analyses consists of 4,484 papers, assigned to 11 categories and 79 subcategories, with 9,644 authors including 7,956 males and 1,688 females. This gender distribution leads to a "baseline" distribution of 17.5% female over the entire dataset.

4.1 Male and Female Authors

Our first analysis consists of a plot of the number of authors by gender over time. Figure 1 shows the absolute number of males and females that had a publication in a certain year, starting with 1971, and ending with 2012. Since our data collection took place toward the end of 2013, the number of publications for 2013 are incomplete, and we therefore removed this year from our time analyses.

Overall, the number of publications grows over time for both males and females. The shape of this graph is influenced by the way the papers were crawled. More popular papers crawled in mid-2013 had a higher probability of appearing in our data set. While this may also be an artifact of the fact that digital publications have been more widely used in recent years (and therefore our collection is biased toward recent times), we believe this trend is also a reflection of a growing number of authors publishing in computer science.

Figure 2 shows the percentage of females among all authors publishing in a 3-year span.[3] We notice that the percentage of female authors increases over time, although we do see some fluctuation between 1980 and present. This includes a large spike around 1980, representing prominent female publications of that time.

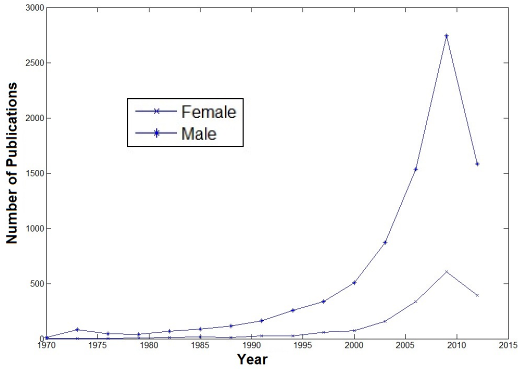

Figure 1: Number of authors by gender by year

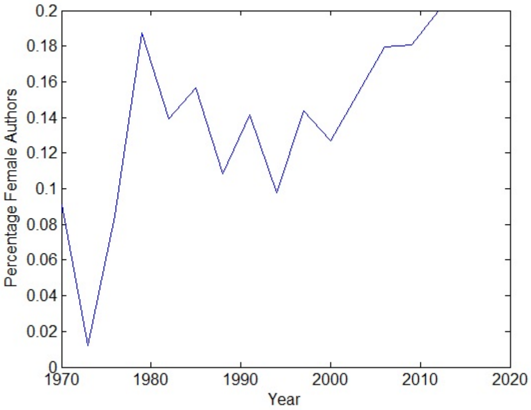

Figure 2: Percentage of female authors by year

4.2 Gender Preferences

To understand the preferences that genders have toward certain categories, we calculate the percentage of female authors in each of the categories used in our paper collection.

Note that although a paper can be assigned to multiple categories, including one or more major categories and zero or more minor categories, for the purpose of these analyses, we consider all the categories equally. We also performed a separate set of analyses that considered only the major category assigned to a paper, but since the findings were very similar, they are not reported here for space reasons.

We first analyze the percentage of female authors on publications belonging to the eleven broad categories listed in Table 1. Table 4 shows the percentage of female authors in each of these categories. The categories for which the difference with respect to the average over the entire collection is significant with $p < 0.05$ are marked with a (*).[4]

Category	Female
Human-centered computing	31.4%*
Applied computing	27.0%*
Information systems	23.0%*
Computing methodologies	21.3%*
Mathematics of computing	20.0%*
Software	19.8%*
Security and privacy	19.6%
Theory	19.1%
Computer systems organization	17.9%
Networks	17.9%
Hardware	16.0%
Micro-average over entire collection	17.5%

Table 4: Percentage of female authors in 11 CSE categories

Human-centered computing and applied computing have the highest percentage of female authors. These categories have a clearly larger fraction of female authors as compared to the other categories, although the highest of the eleven categories still only contains 31% female authors.

We then repeat the same analysis, but this time using the second-level categories. Table 5 shows the percentage of women authors in each of the 79 sub-categories. Note that we only included those sub-categories that had more than 100 publications.

While to our knowledge this is the first time that the gender interests in CSE areas are objectively quantified, the findings are intuitive and inline with previous sociological

[3] For obvious reasons, whenever we calculate percentages, we only need to report the numbers for one gender, as the other gender can be inferred by subtracting from 1. Throughout this paper, percentage calculations refer to female percentages.

[4] We calculate significance using a binomial cumulative distribution function.

research, which suggested that women tend to prefer areas that have a strong applicative component or a clear social aspect (Margolis and Fisher 2003).

Category	Female
Interaction design	33.9%*
Education	32.7%*
Human computer interaction (HCI)	32.0%*
Collaborative and social computing	31.9%*
Law, social and behavioral sciences	29.1%*
Human and societal aspects of security and privacy	28.3%*
Arts and humanities	27.2%*
Electronic commerce	25.0%*
Life and medical sciences	24.7%*
Enterprise computing	24.4%*
World Wide Web	24.1%*
Document management and text processing	23.8%*
Information retrieval	23.6%*
Information systems applications	22.1%*
Information storage systems	21.5%
Computer graphics	21.3%
Artificial intelligence	21.0%
Physical sciences and engineering	20.6%
Software creation and management	20.5%
Network architectures	20.5%
Cryptography	19.4%
Modeling and simulation	19.4%
Computers in other domains	19.3%
Theory and algorithms for application domains	19.3%
Data management systems	19.2%
Information theory	18.8%
Communication hardware, interfaces and storage	18.7%
Mathematical analysis	18.6%
Probability and statistics	18.6%
Software organization and properties	18.3%
Operations research	18.2%
Design and analysis of algorithms	17.9%
Software notations and tools	17.8%
Network properties	17.7%
Machine learning	17.4%
Network types	17.4%
Network services	17.0%
Semantics and reasoning	17.0%
Models of computation	16.5%
Discrete mathematics	16.4%
Network protocols	16.2%*
Architectures	16.2%*
Computational complexity and cryptography	16.1%
Logic	16.0%
Dependable and fault-tolerant systems and networks	15.4%*
Network performance evaluation	15.4%*
Emerging technologies	15.2%
Hardware validation	15.1%
Systems security	14.5%*
Real-time systems	14.4%*
Embedded and cyber-physical systems	14.2%*
Integrated circuits	14.1%*
Symbolic and algebraic manipulation*	14.0%*
Parallel computing methodologies	13.6%
Formal languages and automata theory	13.4%*
Hardware test	13.4%
Very large scale integration design	12.0%*
Robustness	11.8%*
Electronic design automation	11.1%*
Randomness, geometry and discrete structures	10.4%*
Micro-average over entire collection	17.5%

Table 5: Percentage of female authors in 79 second-level CSE categories

4.3 Gender Interests over Time

Using the papers crawled from the digital library, we can also look at publication dates to determine how the number of publications per gender in each category changes over time. The graphs in figures 3 and 4 show number of publications in each category for women and men respectively. The graph for publications by women shows a trend of Applied computing, Computing methodologies, Human-centered computing, and Information systems competing since 1995. More recently Human-centered computing and Applied computing appear higher. Instead, we see that Security and privacy, Hardware, Mathematics of computing, and Computer systems organization have been consistently lower. We also plot the change in the percentage of women authors in different areas over time, as shown in Figure 5.

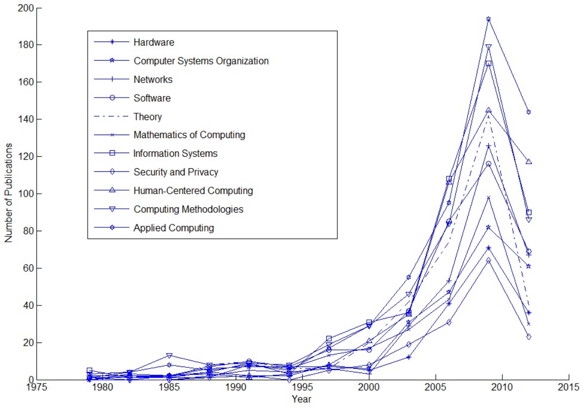

Figure 3: Publications by women in CSE topics by year

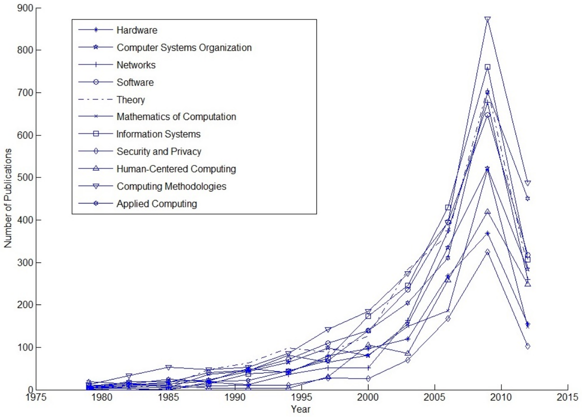

Figure 4: Publications by men in CSE topics by year

5 Conclusions

Our findings suggest that there are areas in CSE that are clearly preferred by women, with a significantly larger fraction of women publishing in these areas as compared to the other areas that are more strongly dominated by men. In particular, we found that Human-centered computing and Applied computing are the two main categories preferred by

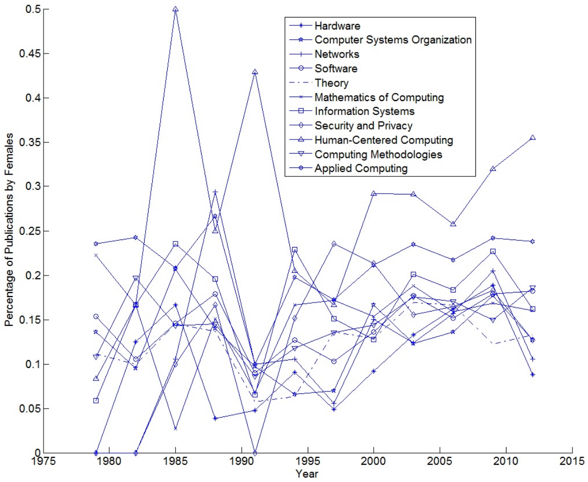

Figure 5: Percentage publications by women in CSE topics by year

women, with a larger than average percentage of women also found in Information systems, Computing methodologies, Mathematics of computing, and Software.

Zooming in, we were also able to identify more specific categories that seem to raise the interest of women, including areas such as Interaction design, Education, Human computer interaction, Social computing, Arts and humanities, and others. Among sub-categories, we were also able to identify areas where the percentage of women is significantly lower than the average, including Randomness, geometry, and discrete structures, VLSI, Hardware tests, Formal languages, Parallel computing, and others.

While there could be a wide variety of factors behind the success of publishing in various CSE categories, we believe our findings are to some degree telling of the interests of men and women in this field. In the future, we plan to diversify our data sources, and also include writings from earlier stages, such as students taking introductory courses in CSE, or students who have not been exposed to CSE.

Acknowledgments

This material is based in part upon work supported by National Science Foundation award #1344257 and by grant #48503 from the John Templeton Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or the John Templeton Foundation.

References

American Society of Engineering Education. 2013. *Profiles of Engineering and Engineering Technology Colleges*. Washington, DC: ASEE.

Angelo, T., and Cross, P. 1993. *Classroom assessment techniques: A handbook for college teachers (2nd ed.)*. San Francisco, CA: Jossey-Bass.

Beauboef, T., and McDowell, P. 2008. Computer science: student myths and misconceptions. *Journal of Computing Sciences in Colleges* 23(6).

Beauboef, T., and Zhang, W. 2011. Where are the women computer science students? *Journal of Computing Sciences in Colleges* 26(4).

Blackburn, R., and Lawrence, J. 1995. *Faculty at work*. Baltimore, Maryland: The John Hopkins University Press.

Burn, H., and Holloway, J. 2006. Why should i care? student motivation in an introductory programming course. In *Annual Conference Proceedings of American Society for Engineering Education*.

Cech, E.; Rubineau, B.; Silbey, S.; and Seron, C. 2011. Professional role confidence and gendered persistence in engineering. *American Sociological Review* 76:641–666.

Claxton, C. S., and Murrell, P. H. 1987. Learning styles: Implications for improving educational practice. ashe-eric higher education report no. 4. Technical report, ASHE, College Station.

Felder, R. 1993. Reaching the second tier: Learning and teaching styles in college science education. *Journal of College Science Teaching* 23(5).

Margolis, J., and Fisher, A. 2003. *Unlocking the Clubhouse: Women in Computing*. MIT Press.

National Academy of Engineering. 2004. *The engineer of 2020: Visions of engineering in the new century*. Washington, DC: National Academy of Engineering.

National Academy of Sciences, National Academy of Engineering, Institute of Medicine. 2007. *Rising above the gathering storm: Energizing and employing America for a brighter economic future*. Washington, D.C.: National Academy Press.

National Science Board. 2004. An emerging and critical problem of the science and engineering labor force: A companion to science and engineering indicators.

National Science Foundation. 2012. Science and engineering indicators.

Newhall, T.; Meeden, L.; Danner, A.; Soni, A.; Ruiz, F.; and Wicentowski, R. 2014. A support program for introductory cs courses that improves student performance and retains students from underrepresented groups.

Seymour, E., and Hewitt, N. 1997. *Talking about leaving*. Boulder, CO: Westview Press.

Spertus, E. 1991. Why are there so few female computer scientists? Technical report, Massachusetts Institute of Technology, Cambridge, MA.

Tang, C.; Ross, K.; Saxena, N.; and Chen, R. 2011. Whats in a name: A study of names, gender inference, and gender behavior in facebook. In Xu, J.; Yu, G.; Zhou, S.; and Unland, R., eds., *Database Systems for Adanced Applications*, volume 6637 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg. 344–356.

Tobias, S. 1990. *Theyre not dumb, theyre different: Stalking the second tier*. Tucson, AZ: Research Corporation.

Werner, L., H. B., and McDowell, C. 2013. Pair-programming helps female computer science students. 4(1).