



# LIWCs the Same, Not the Same: Gendered Linguistic Signals of Performance and Experience in Online STEM Courses

Yiwen Lin<sup>(✉)</sup>, Renzhe Yu<sup>(✉)</sup>, and Nia Dowell<sup>(✉)</sup>

University of California, Irvine, CA 92697, USA  
{yiwenl21,renzhey,dowelln}@uci.edu

**Abstract.** Women are traditionally underrepresented in science, technology, engineering, and mathematics (STEM). While the representation of women in STEM classrooms has grown rapidly in recent years, it remains pedagogically meaningful to understand whether their learning outcomes are achieved in different ways than male students. In this study, we explored this issue through the lens of language in the context of an asynchronous online discussion forum. We applied Linguistic Inquiry and Word Count (LIWC) to examine linguistic features of students' reflective posting in an online chemistry class at a four-year university. Our results suggest that cognitive linguistic features significantly predict the likelihood of passing the course and increases perceived sense of belonging. However, these results only hold true for female students. Pronouns and words relevant to social presence correlate with passing the course in different directions, and this mixed relationship is more polarized among male students. Interestingly, the linguistic features per se do not differ significantly between genders. Overall, our findings provide a more nuanced account of the relationship between linguistic signals of social/cognitive presence and learning outcomes. We conclude with implications for pedagogical interventions and system design to inclusively support learner success in online STEM courses.

**Keywords:** Community of Inquiry · Gender in STEM · Computational linguistics · Linguistic Inquiry and Word Count (LIWC) · Online learning · Higher education

## 1 Introduction

In higher education, introductory courses have been found to have key influence on students' motivation to major in science, technology, engineering, and mathematics (STEM) disciplines [26, 39]. The success in introductory STEM courses is not only determined by academic performance, but also by whether or not students feel supported by the classroom community [19]. While we have

seen an increase in female students' enrollment in STEM disciplines, particularly in online courses [45], more challenges and higher attrition rates were reported among females [3]. It is thus important to understand female students' presence in STEM classrooms and how it affects their learning outcomes and experience. With universities increasingly employing learning management systems and offering STEM classes online, there are more opportunities for learning analytics to offer insights into students' learning processes and for artificial intelligence systems to appropriately scaffold learning behaviors.

Language is a window to learners' social, cognitive, and affective states in learning [8, 10, 13, 42]. The advances of computational linguistics offer a powerful and efficient way to quantify learning behavior at scale [7, 9, 11, 12]. While these methods have been commonly applied to forecast academic achievement and cognitive processes [35, 36], there have been fewer instances that focus on non-cognitive outcomes such as learning experience and social identity [1, 6]. Moreover, prior research suggests that there are gender differences at the socio-linguistic level in computer-mediated communication [5, 29, 32]. But it is less known whether these language patterns are associated with outcomes in a different manner for male and female students. As such, we are interested in exploring whether linguistic characteristics of students' discussion forum posts foretell cognitive and non-cognitive outcomes, and what this means for different genders in the context of online STEM courses.

The contributions of this work are as follows. We extend the understanding of gender differences in STEM learning through the lens of language, illustrating the links between linguistics features in students' reflective posting and performance. Further, by incorporating sense of belonging as an additional outcome measure, we demonstrate different ways language is associated with female and male students' experience in class. Lastly, our research contributes to the emerging research around (gender) equity in personalized and adaptive AIED systems. In the conclusion section, we further discuss the theoretical and practical implications for future research and practices in the AIED community.

## 2 Related Work

The Community of Inquiry (CoI) framework is commonly referenced by works on asynchronous discussion forums. The framework is comprised of three components: cognitive, social, and teaching presence [16]. We primarily focus on the cognitive and social presence in this work. Cognitive presence involves higher-order thinking and constructing meaning through reflection [25]. In the context of our current investigation, the reflection writing assignment highlights two phases of cognitive presence: knowledge integration and resolution. Cognitive presence can be achieved when students link new concepts to past knowledge, and reflect on the application of what they learned in class to real-life scenarios.

Social presence reflects the process when learners interact socially and coordinate efforts with peers [16, 31]. In online learning, social presence is further elaborated as "the ability of participants to identify with the community (e.g.,

course of study)” [15]. Ample evidence from existing literature suggests enhanced academic outcome and educational experience through promoting cognitive and social presence [41]. Prior research also suggests that learning activities promoting social presence may also enhance the learner’s satisfaction and a greater sense of belonging to the online community [2, 17, 37].

There has been extensive research that applies linguistic analysis to reveal social and cognitive presence in online learning. Previously, research has found distinct distributions of psychological categories of words at each level of the cognitive presence in the CoI framework, legitimizing language as a proxy for cognitive engagement [14, 24]. Other research combined natural language processing techniques and behavioral data to establish the connection between linguistic features and engagement to predict learning outcomes [4]. The advances in computational techniques and machine learning models have given rise to automatic identifications of activities in discussion forums that require timely intervention [44]. Researchers have also attempted to translate the CoI coding scheme into a artificial intelligence model to capture cognitive presence [27]. However, it has become increasingly evident that the AIED community should progress towards building automated approaches with an eye on equity and inclusivity in order to appropriately address the issue of “one size fits all” [18].

Previous research suggests that linguistics characteristics in computer-mediated communication differ across gender lines [21, 22]. In the context of STEM learning, a more recent study also found the ability for language to reveal distinct socio-cognitive processes in male and female students’ engagement [29]. With online courses serving as an entryway for female students to pursue STEM disciplines [45], it is important to understand what leads to female students’ performance and learning experiences in introductory STEM courses compared to their male counterparts [14, 30]. Towards that end, we propose the following research questions:

1. What linguistic features of students’ reflective posting are associated with cognitive and non-cognitive outcomes in online STEM courses?
2. Do male and female students exhibit different linguistic features?
3. Do these linguistic features correlate with learning outcomes differently for male and female students?

## 3 Methods

### 3.1 Sample and Data

This study was conducted in a fully online, ten-week introductory chemistry course at four-year university in the United States, with a total of 300 students enrolled. The course was administered in the Canvas learning management system (LMS) and students were required to write a reflection post every week in the discussion forum about the assigned reading for that week. This discussion task accounted for 5% of the final course grade and was organized in small groups of ten students. Each student was randomly assigned to a group at the end of

Week 2 and remained in the same group throughout the course. They could only access the posts written by their group members. Beyond the required posts for course credits, students were free to make additional contribution in the discussion forum. For fair comparison, we focused our text analysis on these original reflection posts.

For the linguistic analysis, we obtained students' discussion posts throughout the course along with their metadata (e.g., timestamp, response relationship). In order to address the first research question, we collected the gradebook data to derive performance measures. For the second research question, a pre- and a post-course survey were sent to measure students' sense of belonging, using a validated Classroom Community Scale [37]. The scale contains ten items on a 5-point Likert scale. For each student, the mean of their valid responses across the ten items was calculated as their sense of belonging. Additionally, we collected students' demographic information and academic history data.

We excluded students who did not post at all throughout the course, leaving a total of 238 students for our final analysis. Among them, 53.6% were female, 42.1% were racial/ethnic minorities (African American, Hispanic and Native American), and 58.6% were first-generation college students. These figures suggest that the class had a fair proportion of traditionally underrepresented student populations in STEM fields, so the findings in this study would be especially meaningful for STEM educators in general.

### 3.2 Linguistic Inquiry and Word Count (LIWC)

Linguistic Inquiry and Word Count (LIWC) is one of the most commonly used dictionary-based tools to evaluate and assess cognitive, social and affective properties in student discourse, as well as educational materials more broadly [34,38]. In the CoI literature, several studies have utilized LIWC to examine the linguistic features associated with social and cognitive presence. In the current research, we focused on a set of LIWC variables that are most representative of cognitive and social presence in students' discussion posts. A brief description of these linguistic features can be found below and in Table 1.

**Table 1.** Summary of LIWC variables in the analysis

Construct	LIWC variables	Example words
Cognitive presence	Analytic, Tone	–
	Cognitive process	Cause, think, should
Social presence	Clout, Authentic	–
	Social process	Friends, talk
	Personal pronouns	I, we, they, you

Among the four composite variables, two of them are used as proxies of cognitive presence. *Analytic* signifies formal and logical language which results

from cognitive processes. *Tone* captures the positive and negative valence in language. Previous research suggests a combination of language valence, pronouns, and cognitive lexicons indicate state of confusion [46]. Academic writing which is less narrative and more cognitively demanding may reside on the negative side of this variable [42]. By contrast, the other two composite variables represent elements relative to social presence. *Clout* is defined as “relative social status, confidence, or leadership displayed through writing” [34]. *Authentic* has been found to signal self-referencing and “humble, vulnerable” positions [33].

The *cognitive process* variable in LIWC includes terms that relate to higher-order thinking and signal cognitive presence [28, 31, 34]. Research has highlighted subcategories of words under this category to demonstrate different phases of cognitive presence [24]. *Social process* includes content words concerning social support and relationships. While this can be a good indicator for social presence in casual contexts, highly social words might conversely suggest off-topicness in formal chemistry reflections. *Personal pronouns* indicate attentional focus and social relationships [35]. Specifically, the use of “I” represents attention drawn to oneself, in contrast to “we”, “you” and “they” which take more “other-oriented” views. Learners who notice and make connections to others’ work are likely to use more other-oriented pronouns [33]. For each student, we computed the average of each LIWC variables across all of their discussion posts to reflect their linguistic experience throughout the term.

### 3.3 Statistical Analysis

We leveraged two models under the framework of generalized linear regressions (GLM) to examine the relationship between linguistic features (all centered and Z-standardized) and students’ cognitive and non-cognitive outcomes. For the cognitive aspect, we used logistic regression to regress the log-odds of passing the course (getting a letter grade of D- or above) on LIWC variables. Note that only 76% of the class passed the course. For the non-cognitive outcome, we used multiple regression, where students’ change in sense of belonging throughout the course was regressed on LIWC variables. In all regression models, students’ background information, including gender, first-generation college status, ethnically underrepresented minority (URM) status and SAT scores, was controlled for, as these variables captured group differences shaped by opportunity gaps prior to their college experience [40]. Also, the four composite LIWC variables were included in separate models from individual LIWC variables (Sect. 3.2) to avoid potential issues of (partial) collinearity.

To compare linguistic features between genders, we used independent t tests to statistically test difference between genders in each of the LIWC variables. Moreover, we reran the previous regression models separately on female and male students, and interpreted the coefficients of LIWC variables to explore potential gender differences in the relationship between linguistic features and student outcomes.

## 4 Results

### 4.1 Linguistic Features and Student Outcomes

Table 2 presents the estimated relationships between LIWC variables and cognitive and non-cognitive outcomes. Note that composite and individual LIWC variables were included in separate regression models. For the cognitive outcome (passing the course), raw instead of exponentiated coefficients from logistic regression models are reported. These estimates show that high cognitive complexity, low social content, negative tones, low social-status language and high frequencies of other-oriented pronouns (we/you/they) are associated with a higher likelihood of passing the course. Reflecting on our construction in Sect. 3.2, these results combined suggest a positive relationship between cognitive presence and cognitive outcome but a more complicated one between social presence and the same outcome. In stark contrast, none of the linguistic features succeeds in predicting students’ change in sense of belonging after taking the course, or the non-cognitive outcome.

**Table 2.** Relationship between LIWC variables and cognitive (passing the course) and non-cognitive (change in sense of belonging) outcomes

	Pass		$\Delta SOB$	
	Coef	SE	Coef	SE
<i>Composite LIWC variables</i>				
Analytic	-.164	(.191)	.079	(.0535)
Clout	-.377*	(.195)	-.0192	(.0615)
Authentic	.0861	(.182)	-.0143	(.061)
Tone	-.473***	(.17)	.0041	(.0684)
<i>Individual LIWC variables</i>				
CogProc	.383*	(.211)	.103	(.0805)
SocProc	-1.04***	(.317)	-.119	(.121)
I	-.057	(.207)	-.07	(.0698)
We	.665*	(.343)	-.0188	(.121)
You	.281	(.186)	-.106	(.0746)
They	.352*	(.197)	.0217	(.0704)

\* p<0.1 \*\* p<0.05 \*\*\* p<0.01

### 4.2 Gender Differences in Linguistic Features

Table 3 presents the summary statistics of LIWC variables for male and female students, respectively. All the statistics were calculated before centering and standardization. The last column reports results from independent t tests to show

**Table 3.** Gender difference in LIWC variables. Format: mean (SD).

	Male	Female	Diff
<i>Composite LIWC variables</i>			
Analytic	83.8 (9.94)	83.8 (10.2)	-.0292
Clout	71.8 (13.4)	72.1 (10.9)	-.325
Authentic	34.6 (15)	36.4 (15.2)	-1.78
Tone	52.1 (16.8)	52.1 (17)	-.0105
<i>Individual LIWC variables</i>			
Cognitive process	11.3 (2.96)	11.3 (2.85)	.04
Social process	6.12 (1.96)	6.16 (1.99)	-.0398
Pronoun: I	.256 (.679)	.237 (.51)	.0191
Pronoun: we	3.35 (1.7)	3.15 (1.53)	.205
Pronoun: you	.176 (.408)	.273 (.431)	-.0973*
Pronoun: they	.558 (.483)	.574 (.514)	-.0166

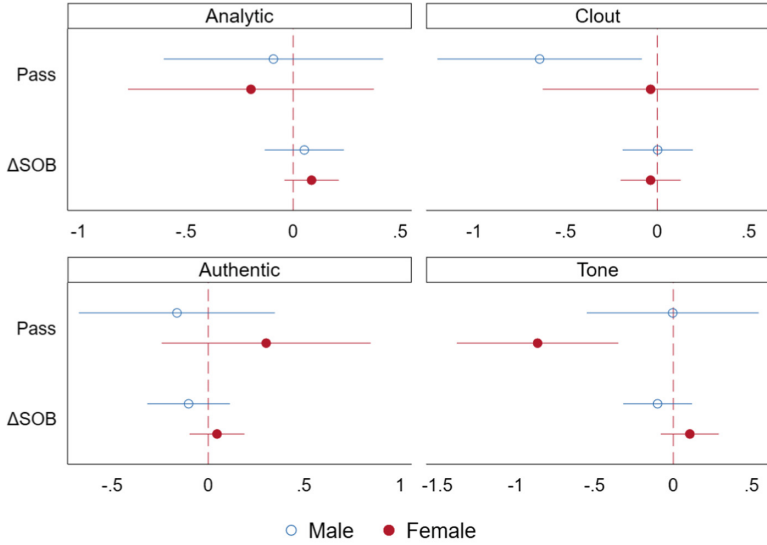
\* p<0.1 \*\* p<0.05 \*\*\* p<0.01

if each variable had a significant gender difference. Contrary to some prior literature [21,22], we did not observe much difference in linguistic features between male and female students. The only differences observed was that male students perceived significantly stronger sense of belonging at the end of the course, and that female students used “you” significantly more in their reflection posts.

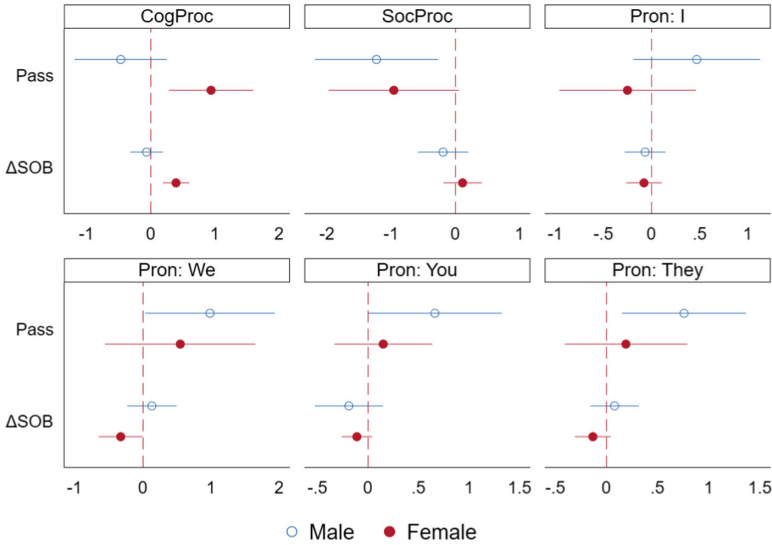
### 4.3 Gender Differences in the Relationship Between Linguistic Features and Student Outcomes

Figure 1 visualizes the estimated coefficients from separate regression models. The visuals depict that the positive relationship between cognitive language and cognitive outcomes is concentrated on female students, evidenced by the significant effects of *tone* (-) and *cognitive process* (+) on the likelihood of passing the course. In contrast, the mixed relationship between social language and cognitive outcomes is more polarized for male students. Specifically, social referencing through other-oriented pronouns (we/you/they) significantly contributes to males’ course outcomes but the use of social words has negative effects on the same outcomes.

While the change in sense of belonging is not correlated with any LIWC variables in the overall model, there are some significant relationships among female students. More cognitive language use predicts an increase in women’s perceived classroom community, whereas other-oriented pronouns exhibit negative associations.



(a) Composite LIWC variables



(b) Individual LIWC variables

**Fig. 1.** Gender differences in the estimated relationship (regression coefficients) between LIWC variables and cognitive (passing the course) and non-cognitive (change in sense of belonging) outcomes



## 5 Discussion

In this study, we investigated the relationships between linguistic features of students' reflective posting and student outcomes in an introductory online chemistry class. We further examined the gender differences in these linguistic features, and in the way they associated with outcomes. From our results, the strong positive relationship between cognitive language use and course performance for female students suggests that there might be an underlying need for female students to demonstrate cognitive engagement through language to achieve better outcomes. Additionally, the positive correlation between cognitive language and increased sense of belonging indicate that females are more likely to derive a sense of belonging from making intellectual contributions to the discussion forum. This might imply that cognitive language can improve learning experience and shape STEM identity more for female students than for male students.

The overall negative relationship between social language and passing the course may suggest that being on-topic is an important indicator of grades [43]. A reflection post with too many social signal words could mean a deviation from core content, leading to lower performance on tests. Regarding the use of pronouns, "we" was associated with decreased perceived sense of belonging for female students, which was somewhat surprising. While we expected that the use of an inclusive pronouns such as "we" would create a greater sense of community, this result shows the opposite. This counter-intuitive relationship might be accounted for by group factors. For instance, if a female student is the only person in her discussion group who engages in deep reflection, she may feel disconnected. A weaker sense of belonging may therefore be triggered by using "we" when the personal and group identity do not align. Due to the scope of our analysis, the current study did not take into account of group-level influence, but this remains an important direction for future work.

## 6 Conclusion

The naturally occurring educational discourse data within online learning platforms presents a golden opportunity for the AIED community to advance the understanding of cognitive and social processes in STEM learning and enables new kinds of personalized interventions focused on increasing inclusivity and equity [20]. Towards these ends, there are several key obstacles including limited analytical approaches to handling the scale of such data and substantive data-driven knowledge that can direct us to cultivate more equitable, respectful, and diverse environments that meaningfully engage all learners. In this context, our findings present some theoretical and practical implications for the AIED community.

For starters, our results alert that transferring and interpreting learner behavior across different types of online environments (i.e., MOOCs versus accredited university classes) or across academic disciplines require careful considerations.

One might assume that increased social presence in asynchronous discussion forums reflected by social language use would benefit learning. Yet the opposite result in the context of this chemistry course suggests that discussing non-academic content may also be irrelevant and undesirable in a formally structured discussion environment. Consequently, contextual information including classroom community and course delivery needs to be considered when deploying AIED applications focused on linguistic analytics. More nuanced considerations should also be given to applying theoretical models to online environments. For the same results above, it is also likely that social presence built upon knowledge construction is more valuable to learners' sense of belonging than that upon shared personal interests. Knowing this differentiation can be particularly informative for designing strategies to reduce the attrition rates of female students in STEM subjects.

Finally, our findings shed light on the emerging discourse around fairness and equity issues in student models [23,47]. Mining educational data should not be left without considerations for equity and inclusivity for different student populations. In our case, although the linguistics features appeared to be indistinguishable for male and female students overall, they were in fact associated with learning outcomes differently at a deeper level. We further highlight concerns about making instructional decisions based on the analysis for an entire student body. Such an approach, as we have found, might inadvertently discount the disparate impact on gender subgroups. Future development of automated analytic tools and machine learning models used to monitor learners' discussion forums activities should thus aim to recognize gender differences in order to close gender gaps in STEM education.

**Acknowledgement.** This paper is based upon work supported by the National Science Foundation (Grant Number 1535300). We thank Peter McPartlan, Qiujie Li and Teomara Rutherford for providing access to the course information and datasets.

## References

1. Abe, J.A.A.: Big five, linguistic styles, and successful online learning. *Internet High. Educ.* **45**, 100724 (2020)
2. Arbaugh, J., Benbunan-Finch, R.: An investigation of epistemological and social dimensions of teaching in online learning environments. *Acad. Manag. Learn. Educ.* **5**(4), 435–447 (2006)
3. Chen, X.: Stem attrition: College students' paths into and out of stem fields (nces 2014–001). Technical report (2013)
4. Crossley, S., Mcnamara, D.S., Paquette, L., Baker, R.S., Dascalu, M.: Combining click-stream data with NLP tools to better understand MOOC completion. In: *ACM International Conference Proceeding Series*, vol. 25–29 April 2016, pp. 6–14. Association for Computing Machinery (2016)
5. Dowell, N., Lin, Y., Godfrey, A., Brooks, C.: Promoting inclusivity through time-dynamic discourse analysis in digitally-mediated collaborative learning. In: Isotani, S., Millán, E., Ogan, A., Hastings, P., McLaren, B., Luckin, R. (eds.) *AIED 2019. LNCS (LNAI)*, vol. 11625, pp. 207–219. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-23204-7\\_18](https://doi.org/10.1007/978-3-030-23204-7_18)

6. Dowell, N., Lin, Y., Godfrey, A., Brooks, C.: Exploring the relationship between emergent sociocognitive roles, collaborative problem-solving skills and outcomes: a group communication analysis. *J. Learn. Anal.* **7**(1), 38–57 (2020)
7. Dowell, N., Poquet, O., Brooks, C.: Applying group communication analysis to educational discourse interactions at scale. *International Society of the Learning Sciences* (2018)
8. Dowell, N.M., Graesser, A.C., Cai, Z.: Language and discourse analysis with co-metrix: Applications from educational material to learning environments at scale. *J. Learn. Anal.* **3**(3), 72–95 (2016)
9. Dowell, N.M., et al.: Modeling learners' social centrality and performance through language and discourse. In: *International Educational Data Mining Society* (2015)
10. Dowell, N.M.M., Graesser, A.C.: Modeling learners' cognitive, affective, and social processes through language and discourse. *J. Learn. Anal.* **1**(3), 183–186 (2014)
11. Dowell, N.M., Brooks, C., Kovanović, V., Joksimović, S., Gašević, D.: The changing patterns of MOOC discourse. In: *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale*, pp. 283–286 (2017)
12. Dowell, N.M.M., Nixon, T.M., Graesser, A.C.: Group communication analysis: a computational linguistics approach for detecting sociocognitive roles in multiparty interactions. *Behav. Res. Methods* **51**(3), 1007–1041 (2018). <https://doi.org/10.3758/s13428-018-1102-z>
13. D'Mello, S.K., Dowell, N., Graesser, A.: Unimodal and multimodal human perception of naturalistic non-basic affective states during human-computer interactions. *IEEE Trans. Affect. Comput.* **4**(4), 452–465 (2013)
14. Fesler, L., Dee, T., Baker, R., Evans, B.: Text as data methods for education research. *J. Res. Educ. Eff.* **12**(4), 707–727 (2019)
15. Garrison, D.R.: Communities of inquiry in online learning. In: *Encyclopedia of Distance Learning*, 2nd edn., pp. 352–355. IGI Global (2009)
16. Garrison, D.R., Anderson, T., Archer, W.: Critical thinking, cognitive presence, and computer conferencing in distance education. *Int. J. Phytorem.* **21**(1), 7–23 (2001)
17. Garrison, D.R., Arbaugh, J.B.: Researching the community of inquiry framework: review, issues, and future directions. *Internet High. Educ.* **10**(3), 157–172 (2007)
18. Gašević, D., Dawson, S., Rogers, T., Gasevic, D.: Learning analytics should not promote one size fits all: the effects of instructional conditions in predicting academic success. *Internet High. Educ.* **28**, 68–84 (2016)
19. Gasiewski, J.A., Eagan, M.K., Garcia, G.A., Hurtado, S., Chang, M.J.: From gate-keeping to engagement: a multicontextual, mixed method study of student academic engagement in introductory stem courses. *Res. High. Educ.* **53**(2), 229–261 (2012). <https://doi.org/10.1007/s11162-011-9247-y>
20. Goldstone, R.L., Lupyan, G.: Discovering psychological principles by mining naturally occurring data sets. *Top. Cogn. Sci.* **8**(3), 548–568 (2016)
21. Guiller, J., Durndell, A.: Students' linguistic behaviour in online discussion groups: does gender matter? *Comput. Hum. Behav.* **23**(5), 2240–2255 (2007)
22. Herring, S.C.: Gender differences in CMC: findings and implications. *Comput. Prof. Soc. Responsib. J.* **18**(1) (2000). <http://archive.cpsr.net/publications/newsletters/issues/2000/winter2000/herring.html>. Accessed 23 Jan 2020
23. Hutt, S., Gardner, M., Duckworth, A.L., D'Mello, S.K.: Evaluating fairness and generalizability in models predicting on-time graduation from college applications. In: *The 12th International Conference on Educational Data Mining (EDM)*, Montréal, Canada, pp. 79–88 (2019)

24. Joksimovic, S., Gasevic, D., Kovanovic, V., Adesope, O., Hatala, M.: Psychological characteristics in cognitive presence of communities of inquiry: a linguistic analysis of online discussions. *Internet High. Educ.* **22**, 1–10 (2014)
25. Kilis, S., Yıldırım, Z.: Investigation of community of inquiry framework in regard to self-regulation, metacognition and motivation. *Comput. Educ.* **126**, 53–64 (2018)
26. Koenig, K., Schen, M., Edwards, M., Bao, L.: Addressing stem retention through a scientific thought and methods course. *J. Coll. Sci. Teach.* **41**(4), 23–29 (2012)
27. Kovanovic, V., Joksimovic, S., Gasevic, D., Hatala, M.: Automated Cognitive Presence Detection in Online Discussion Transcripts (2014). <http://nlp.stanford.edu/software/corenlp.shtml>
28. Kramer, I.M., Kusurkar, R.A.: Science-writing in the blogosphere as a tool to promote autonomous motivation in education. *Internet High. Educ.* **35**, 48–62 (2017)
29. Lin, Y., Dowell, N., Godfrey, A., Choi, H., Brooks, C.: Modeling gender dynamics in intra and interpersonal interactions during online collaborative learning. In: Proceedings of the 9th International Conference on Learning Analytics & Knowledge, pp. 431–435 (2019)
30. Marie Jackson, S., Marie, S.: The influence of implicit and explicit gender bias on grading, and the effectiveness of rubrics for reducing bias repository citation. Technical report. [https://corescholar.libraries.wright.edu/etd\\_all/1529](https://corescholar.libraries.wright.edu/etd_all/1529)
31. Moore, R.L., Oliver, K.M., Wang, C.: Setting the pace: examining cognitive processing in MOOC discussion forums with automatic text analysis. *Interact. Learn. Environ.* **27**(5–6), 655–669 (2019)
32. Nguyen, D., Doğruöz, A.S., Rosé, C.P., de Jong, F.: Computational sociolinguistics: a survey. *Comput. Linguist.* **42**(3), 537–593 (2016)
33. Oliver, K.M., Houchins, J.K., Moore, R.L., et al.: Informing makerspace outcomes through a linguistic analysis of written and video-recorded project assessments. *Int. J. Sci. Math. Educ.* (2020). <https://doi.org/10.1007/s10763-020-10060-2>
34. Pennebaker, J.W., Boyd, R.L., Jordan, K., Blackburn, K.: The development and psychometric properties of LIWC2015. Technical report (2015)
35. Pennebaker, J.W., Chung, C.K., Frazee, J., Lavergne, G.M., Beaver, D.I.: When small words foretell academic success: the case of college admissions essays. *PLoS One* **9**(12), e115844 (2014)
36. Robinson, C., Yeomans, M., Reich, J., Hulleman, C., Gehlbach, H.: Forecasting student achievement in MOOCs with natural language processing. In: Proceedings of the Sixth International Conference on Learning Analytics & Knowledge, pp. 383–387 (2016)
37. Rovai, A.P.: Development of an instrument to measure classroom community. *Internet High. Educ.* **5**(3), 197–211 (2002)
38. Sell, J., Farreras, I.G.: Liwc-ing at a century of introductory college textbooks: have the sentiments changed? *Procedia Comput. Sci.* **118**, 108–112 (2017)
39. Seymour, E., Hewitt, N.M.: *Talking About Leaving*. Westview Press, Boulder (1997)
40. Shapiro, D., Dundar, A., Huie, F., Wakhungu, P., Bhimdiwala, A., Wilson, S.: Completing college: a state-level view of student completion rates (signature report no. 16a). Technical report, National Student Clearinghouse Research Center, Herndon, VA (2019)
41. Swan, K., Matthews, D., Bogle, L., Boles, E., Day, S.: Linking online course design and implementation to learning outcomes: a design experiment. *Internet High. Educ.* **15**(2), 81–88 (2012)

42. Tausczik, Y.R., Pennebaker, J.W.: The psychological meaning of words: LIWC and computerized text analysis methods. *J. Lang. Soc. Psychol.* **29**(1), 24–54 (2010)
43. Wise, A.F., Cui, Y.: Unpacking the relationship between discussion forum participation and learning in MOOCs. In: *Proceedings of the 8th International Conference on Learning Analytics and Knowledge - LAK 2018*, pp. 330–339. ACM Press, New York (2018)
44. Wise, A.F., Cui, Y., Jin, W.Q., Vytasek, J.: Mining for gold: Identifying content-related MOOC discussion threads across domains through linguistic modeling. *Internet High. Educ.* **32**, 11–28 (2017)
45. Wladis, C., Hachey, A.C., Conway, K.: Which STEM majors enroll in online courses, and why should we care? The impact of ethnicity, gender, and non-traditional student characteristics. *Comput. Educ.* **87**, 285–308 (2015)
46. Yang, J.C., Quadir, B., Chen, N.S., Miao, Q.: Effects of online presence on learning performance in a blog-based online course. *Internet High. Educ.* **30**, 11–20 (2016)
47. Yu, R., Li, Q., Fischer, C., Doroudi, S., Xu, D.: Towards accurate and fair prediction of college success: evaluating different sources of student data. In: *Proceedings of the 13th International Conference on Educational Data Mining (EDM 2020)* (2020)