How Gender Cues in Educational Video Impact Participation and Retention

Christopher Brooks, School of Information, University of Michigan, brooksch@umich.edu
Josh Gardner, School of Information, University of Michigan, jpgard@umich.edu
Kaifeng Chen, School of Information, University of Michigan, chenkf@umich.edu

Abstract This work describes a large-scale randomized experiment in an introductory data science MOOC on the Coursera platform. The experiment subtly altered gendered situational cues in course videos, placing either male or female data scientists in video backgrounds and using matching male or female aides for tutorial videos. The experiment explores whether prior work on ambient cues would generalize beyond a physical classroom environment and into an online environment at scale. We find strong evidence that the female condition induces strong positive effects on overall course activity and discussion posting behavior by female students, but also strong evidence of a smaller negative effect on these outcomes for male students. We find no effect on students' persistence through the course. This experiment suggests that subtle personalized alterations of educational environments can influence students' engagement patterns in large-scale digital learning environments, but that gendered interventions may negatively impact some students.

Introduction

MOOCs offer the potential for millions of learners around the globe to access high-quality educational content at scale. Despite reaching incredible numbers of students (1), MOOCs have seen far less of the experimental learning science research that has evaluated traditional educational environments such as in-person primary and secondary education. As a result, little is known about how course environments and mechanisms of delivery within the learning environment can influence MOOC participants. This is a particularly noteworthy literature gap because of the heterogeneity of most MOOC environments relative to in-person educational settings: students can learn from, interact with, and passively observe other learners with vastly different cultures, backgrounds, and motivations (Kizilcec and Brooks 2017). This background can influence how learners participate in this educational phenomena (Kizilcec and Cohen 2017).

In this work, we examine the effect of situational cues relating to gender in an introductory data science MOOC. This course, which has had awarded nearly ten thousand certificates and engaged nearly one hundred thousand learners in its first year, has significant gender imbalance, with roughly 84% of the learners being male. In addition, persistence in the course is lower for women than men (see Figure 1).

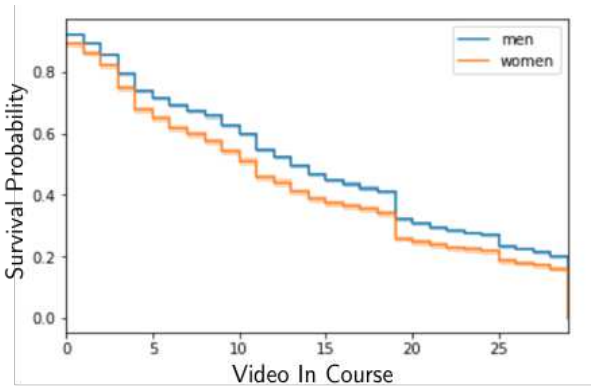

Figure 1. Kaplan-Meier survival curves for men (blue, top) and women (orange, bottom) enrollees over time (expressed as the order of the lecture in the course). Note the gap between curves (significant at p≈0.0000 using a log-rank test) demonstrates the lack of retention of women, even as early as the first lecture in the course.

The current work is guided by prior research by Cheryan et al. (2009) which evaluated how stereotypical cues within the learning environment, even those unrelated to content (i.e., posters and objects in the classroom, such as Star Wars posters or video games), can affect female students' sense of belonging when considering enrolling in the discipline of Computer Science. Cheryan et al. conducted a randomized in-person experiment (n = 39, women = 22), where stereotypically masculine or neutral cues were crowdsourced from

outside of the subject pool and placed in physical classrooms before subjects were introduced. The subjects were told they were participating in a survey on interest in technical jobs and internships. Cheryan et al. find that female students who participated in the stereotypically masculine classroom condition were significantly less interested in computer science than were men, but that there was no gender difference with respect to interest in computer science in the neutral classroom. Importantly, under the neutral condition there was (i) no statistically significant impact on male students, and (ii) no observable gender performance difference on a short STEM-based assessment item.

Other research has demonstrated effects related to stereotype threat -- the concern that others will judge one negatively due to a stereotype that exists about one's group -- for female students in engineering environments. In these environments, stereotypes can negatively affects female students' performance in engineering coursework, but Bell et al. (2003) showed that making contextual changes which reduce stereotype threat (i.e., changing the framing of a test from being a diagnostic of ability to being non-diagnostic and not producing gender differences) can eliminate performance gaps.

We are particularly interested in understanding scalable mechanisms to address these issues in large online courses. Ideally, such mechanisms could be automatically delivered based on user models of the learner, customizing the environment to best accommodate their success. With MOOCs being particularly reliant on video-based instruction, and with these videos being such a salient aspect of the learner experience, we focused our experimentation with potential customizations of the video itself, altering elements of the video with respect to gender representation (described more fully in Section 2: Experiment). These alterations resulted in the creation of two conditions, one situational cues aimed at supporting women (the *female condition*) and one with situational cues aimed to support men (the *male condition*). We then pre-registered (2) the following hypotheses:

> H1: That there will be higher retention rates among women in the female condition than of women in the male condition.
>
> H3: That there will be higher ... (c) forum participation ... of women in the female condition than of women in the male condition.
>
> H4: That there will be no difference in the rate of participation of males between the two conditions.
>
> (Brooks & Gardner, 2017)

Experiment

The experiment took place in a large data science MOOC hosted on the Coursera platform. Students using the platform must provide their name and email address to Coursera, and optionally fill out a demographics survey. Coursera records the gender for those who fill it out in the demographics survey, and runs an inferencing process (3) to determine gender of participants by name otherwise. Learners were randomly assigned to either the female (n = 23,365, male students = 18,482) or male conditions (n = 23,287, male students = 18,478). The course consists of a total of 29 lectures and several assignments. Learners can sign up for free (auditing; no graded assessments available), paid (full course including assessments, either at a flat or monthly rate), or for financial aid (experience the same as paid learners). Upon successful completion of either the paid or financial aid experience, learners are provided with a certificate of achievement. Data collection for the current analysis was conducted over a ten month period, with new cohorts beginning every two weeks.

Each condition differs only in two ways. First, all of the videos which include the instructor are shown with either men (Figure 2a) or women (Figure 2b) working in the background. Learners were introduced in the first video to this space as an educational technology incubator where data scientists were employed. We intentionally chose individuals to portray data scientists who exhibited physical masculine (e.g. the bearded individual in Figure 2a) or feminine (e.g. the woman wearing a skirt in figure 2b) traits. Actual video was composited in post-processing, with the instructor added via green screen chroma keying. Secondly, a tutorial assistant who delivered four lectures over the length of the course we either a man (Figure 2c) or a woman (Figure 2d) and were both introduced using a gendered name and wore stereotypical masculine or feminine dress. As many learners search for and connect with MOOC instructors on social media, and the instructional content is long and would be expensive to refilm, it was deemed intractable to change the video of the instructor directly.

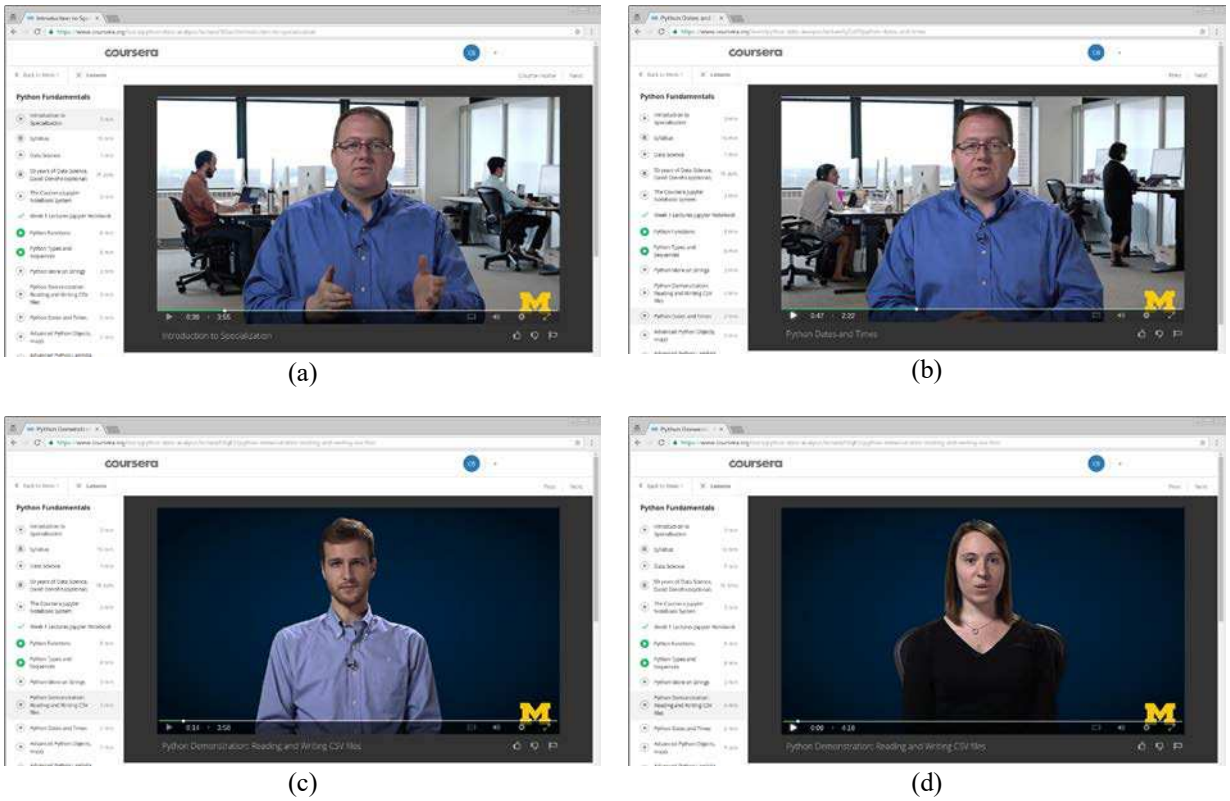

Figure 2. Male condition (left) and female condition (right). The top row shows the instructional video setting for each condition; the bottom row shows the assisted tutorial settings for each condition.

Results

Bayesian analysis

We evaluated the results primarily through the use of Bayesian data analysis techniques. Bayesian methods have much to offer learning science researchers, where evaluating experimental data under uncertainty about the underlying effects is common. In contrast to frequentist methods, which often make strong assumptions (as, for example, a *t*-test might assume normality across groups, or an ordinary least squares regression model might assume normally-distributed errors; both tests provide inferences conditional on the assumption that the null hypothesis is correct), Bayesian methods have the ability to incorporate uncertainty about these assumptions directly into the model through hierarchical modeling (J. K. Kruschke 2013). The Bayesian approach is particularly appropriate for the analysis of large-scale randomized experiments (J. K. Kruschke and Liddell 2015), such as those involving data from MOOCs. First, when the sample size is sufficiently large, significance tests of a null hypothesis of equivalence will always reject a null hypothesis of equivalence in the presence of any observed difference, no matter how small, because *p*-values are affected by both sample size and effect size (Wasserstein and Lazar 2016). Bayesian approaches differentiate between sample size and effect size (and estimate effects more precisely, but not necessarily with greater "significance" or magnitude, as *N* increases) (J. K. Kruschke and Liddell 2015). Second, Bayesian techniques support *probabilistic* statements about treatment effects, given the data. This is in contrast to the black-and-white conclusions usually drawn using null hypothesis significance testing methods, which provide little additional information about the uncertainty of the effect being observed, and which provide these conclusions conditional on the assumed correctness of the null hypothesis (typically, this is a null hypothesis that no effect exists, or that the difference between conditions is precisely zero). It has been widely noted that a null hypothesis of zero effect is often unrealistic, and a test which uses this hypothesis provides no evidence for whether the null hypothesis itself may be true (Cohen 1994). Finally, Bayesian estimates are not subject to frequentist concerns about multiple comparisons, because Bayesian analysis is only concerned with the posterior distribution based on the actual data -- not with

hypothetical unobserved data -- and because the hierarchical structure of Bayesian models imposes data-driven shrinkage on estimates under uncertainty (Gelman et al. 2012, Kruschke and Liddell 2015).

Retention (H1)

As a measure of retention throughout the course we calculated furthest video in the course a learner watch (regardless of order). We model this as an *ordinal* -- in contrast to metric -- outcome, for several reasons: different course weeks have different numbers of videos; some videos are required content and others are optional; and videos are different lengths (ranging from less than two minutes to 10 minutes or more). We adapt an ordinal logistic regression model (J. Kruschke 2014) to model the probability distribution of a learner in each condition dropping out at each of 29 videos in the course. The results of this model are shown in Figure 3. In particular, Figure 3c shows a comparison of the Bayesian Credible Intervals for the two groups with 95% credible intervals and there is little, if any, discernible effect on learners' penetration into the course as measured by their maximum video watched. Indeed, for the female students shown here, the probabilities of dropout at each video are nearly identical across each condition. As the 95% credible intervals overlap for each video we conclude there is no difference for retention of women between conditions, and H1 does not hold.

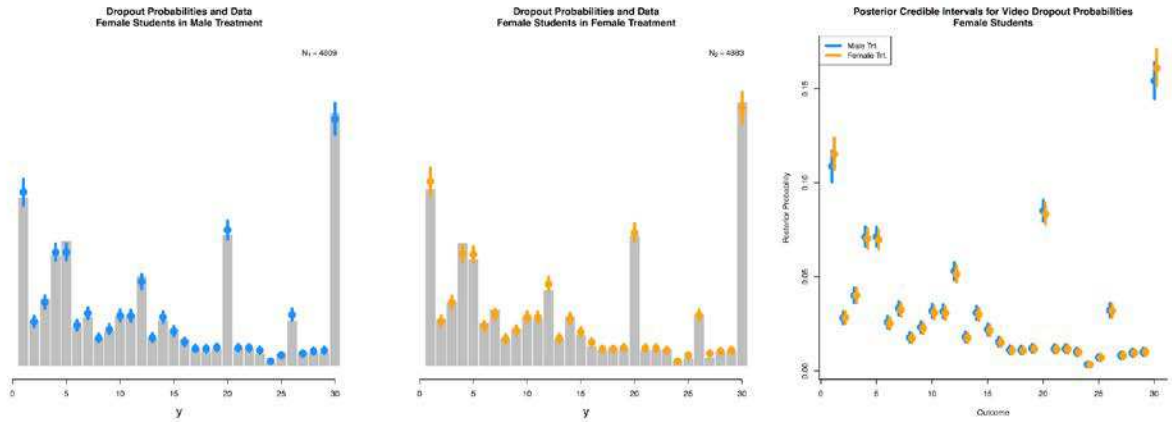

Figure 3. Actual (bars) vs. predicted (intervals) dropout probabilities for female students in the female (left) and male (center) conditions, and (right) a comparison of the Bayesian Credible Intervals for the two groups, which shows little evidence of an effect on the outcome evaluated.

Forum posting (H3c, H4c)

In addition to instructional videos, discussion forums are a primary component of courses, and are notable because participation in the discussion fora is entirely voluntary. We were therefore interested in whether condition assignments had an effect on the likelihood or frequency of users' engagement with course discussion forums. Because this analysis models a count-based outcome, we used a zero-inflated hierarchical Poisson model, which incorporated a latent variable to model learners with zero posts. This model allowed us to estimate (a) a latent variable indicator for whether a learners in each group would participate in the course (whether they would have a non-zero number of clickstream entries), and (b) for forum posters, an estimate of their expected number of posts (this is operationalized as the lambda parameter of a Poisson model, which is both the mean and variance of the distribution). The model specification is:

$$
\begin{aligned}
Y_i|\delta_i = 1, T_i = 1 &\sim Poisson(\lambda_1) \\
Y_i|\delta_i = 1, T_i = 0 &\sim Poisson(\lambda_0) \\
\delta_i|T_i = 1 &\sim Bernoulli(\pi_1) \\
\delta_i|T_i = 0 &\sim Bernoulli(\pi_0)
\end{aligned}
$$

$T_i$is an indicator for being in the female condition,$\delta_i$is a latent variable indicating whether the outcome of interest has a non-zero value, and $Y_i$is the outcome of interest (which is zero when $\delta_i = 0$). Standard noninformative priors were used for the hyperparameters, with $\pi \sim Uniform(0,1)$and $\lambda \sim Gamma(0.001, 0.001)$.

This model estimated no difference in the probability of posting for students of either gender, in either treatment branch; the 95% Bayesian Credible Interval for $\pi_1 - \pi_0$included zero for each gender, with the probability of posting for each group $\pi_1 \approx \pi_0 \approx 3.5\%$. However, for students who did post in the discussion forums, the female branch induced an increase of 1.78 posts per student, with the estimated mean for females in the female branch at 6.28 versus 4.62 for those in the male branch; the probability of the true difference being positive (for the female branch) for this outcome was asymptotic to one based on 10,000 Markov chain Monte Carlo (MCMC) samples from the posterior. We observed a smaller, negative effect of the female condition for male students. The estimated mean of male students in the female condition who posted in the discussion fora was 5.67 posts per student, versus 6.31 posts for those in the male condition, a *decrease* of 0.64 posts.

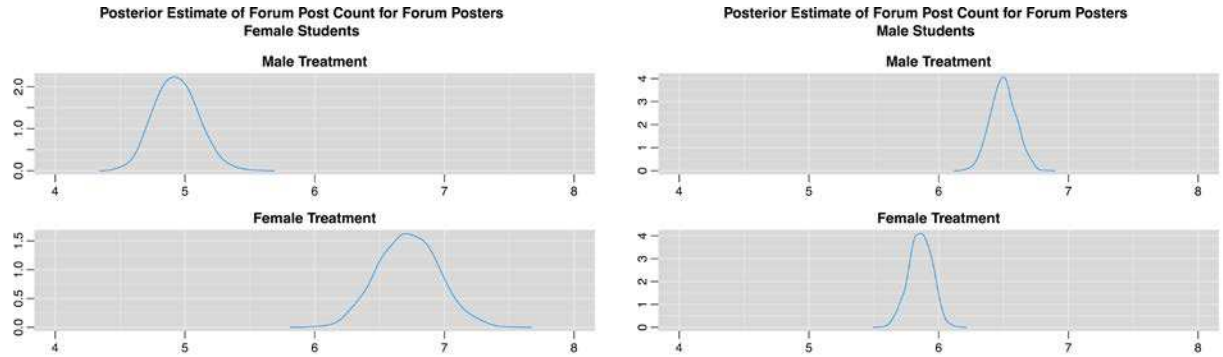

Figure 4. Posterior distributions of the number of discussion forum posts for female (left) and male (right) students in each condition who utilize the discussion fora. Note that there was no difference in the probability of posting for either group, according to the latent variable model used.

Engagement and activity (exploratory)

Another primary outcome of interest for this experiment was female students' engagement with course content. As such, we evaluated the treatment effect for both female and male students separately, comparing the effect of different treatments within each gender (each of our analyses follows this approach). To evaluate students' overall engagement with the course, we extracted the total number of clickstream entries for each student (each clickstream entry constitutes a request to the Coursera server, such as a page load, video view, or click). We employed a latent variable model with an identical structure as described in the previous section to assess the impact of each treatment branch on each gender and treatment group's propensity to engage with the course platform, and the number of clickstream events for students who did engage with the course.

Motivated by the forum analyses, we explored the clickstream data in an effort to build more theory around engagement. Specifically, we modeled the total number of entries in the clickstream logs for a user (excluding video heartbeat entries, which indicate when a video is loaded) as a coarse-grained measure of engagement. Results from this analysis are shown in the posterior density plots in Figure 4. For female learners, the female condition generated an average of 37.6 additional clickstream entries for students who participated in the course (those with nonzero clickstream counts), with a posterior estimated average of 480.0 clickstream events per active female student in the female condition, versus 442.3 for those in the male condition (an increase of 8.5%). Paralleling the effects observed in the forum analysis above, for male learners, the effect was in the opposite direction, but much smaller: the female condition produced an average *decrease* of 11.6 clickstream entries for students in this branch, relative to those in the male branch. Posterior estimated average clickstream events were 496.2 for male students in the female branch, versus 507.8 for those in the male branch (a decrease of 2.3%). There was no difference in the probability of students actively participating in the course in either treatment branch; we would expect this, as the treatment would not be visible to students who never accessed the course. Interactions of male students on average produced more clickstream events in the platform across all conditions.

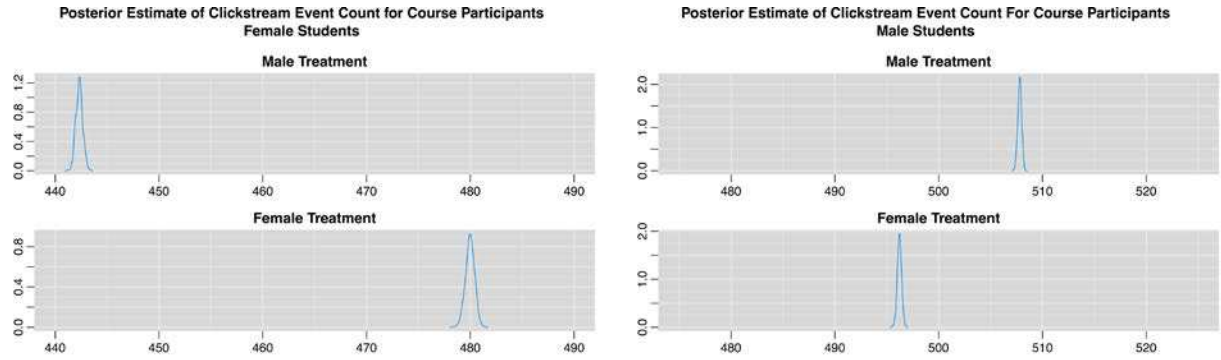

Figure 5. Posterior distributions of the number of clickstream events for female (left) and male (right) students in each condition.

Discussion

The issue of inclusivity in online learning environments is poorly understood, and most of the existing literature is conducted in traditional face-to-face educational environments. Yet a prototypical technology for the area of Artificial Intelligence in Education (AIED), for instance, is the adaptive and personalized learning system. MOOCs are one of the contexts in which personalized learning stands to have a significant impact, due to the sheer numbers of learners engaging in those systems. However, there has been limited experimentation with adaptive systems in MOOCs to date, even as studies show interventions and achievement vary by social group (Kizilcec and Cohen 2017; Kizelcec, Saltarelli, Reich and Cohen, 2017). In this work, we have aimed to impact the intersection of these fields with the learning sciences, and provide new evidence toward understanding the impact of situational cues in large scale online technical courses.

More specifically, we have demonstrated that stereotypically masculine or feminine cues in learning environments are not universally more or less welcoming to students of a gender matching the cues, and that the magnitude and direction of their effect varies by subjects' gender. This adds additional context to the results of the highly cited work of Cheryan et al. (2009), particularly in the case of Massive Open Online Courses. The distinctions between our experimental setup are important; whereas Cheryan et al. studied a small number of residential students at Stanford and their intent to enroll in Computer Science when introduced to the discipline in a crowdsourced, stereotypically masculine environment, we studied a large number of global learners and their activities (retention, forum posting, and overall engagement) while engaging in learning in an online environment with video-based situational gender cues. The differences in the experimental parameters, and our results, are significant, and clearly more work is needed to understand the impact of situational gender cues on student behavior. Our work here suggests that while retention is not affected by these cues, engagement in forums and overall course activity is, for both women and men.

The question of the underlying mechanisms related to environmental cues in video content, and why these mechanisms might or might not affect learners, is one which we have not examined here. Future work with our existing dataset, looking both at the qualitative analysis of learner perceptions as well as analysis of forum discussions, may reveal underlying mechanisms which mediate posting behavior. Regardless, the posteriors for both discussion forum posts (Figure 4) and clickstream entries (Figure 5) are strongly suggestive of an effect which varies in both magnitude and direction by gender, and the analysis provided for H1 suggests there is no strong effect on retention. Future studies varying the nature and frequency of gender-based environmental cues in diverse content and pedagogical environments are warranted in order to generalise this finding. Further, the question as to why the two populations showed different levels of sensitivity to the environmental cues is one which we have not explored, though work by Rudman and Goodwin (2004) has demonstrated that in-group biases are stronger for women in some situations. We note that despite having a greater impact on women (an increase of 1.78 posts per student, versus 0.64 posts for men), if the female treatment condition were deployed to all students at scale there would be a net loss of engagement due to the high gender imbalance in the course (84% male).

This implications of this work go beyond the immediate issue of inclusivity with gender, and begs for further experimentation on *personalization of video at scale*. Modern computing power, storage, and bandwidth makes real-time composition of video a tractable way to personalize a learning environment. Video makes up the backbone of MOOC experiences, and is a high value artifact which stems from the instructional and media

design activity which goes into offering these courses. Instead of focusing on producing "the best video", we argue that not only are there are many different videos which may be appropriate for different audiences using visual cues alone, but that we know very little about how such cues would impact learners at present. While our work has looked specifically at gender cues in video environments, it is not unreasonable to consider that there may be other environmental cues which have priming effects on learners. While it is not currently feasible to replace the main instructors (at least in scaled production environments) of MOOC educational video, we point to the vision of Stephenson (1995) in the fictional *The Diamond Age: Or, A Young Lady's Illustrated Primer* whereby the educational content itself was dynamically assembled based on the learner, the context, and the interpretations of the actions of live actors, facilitated in party by software systems similar to the current gig-economy. In this world, the ability to rapidly experiment with and learn the impacts of diverse sets of environmental cues would be possible, and through innovation in personalized video we believe that the areas of the learning sciences, learning at scale, and artificial intelligence in education are well positioned to provide both impact from and evidence on the effects of environmental cues.

Endnotes

(1) The Coursera platform, used in the current experiment, now boasts 29 million unique learners alone.

(2) Additional hypotheses which focused on regional sub-populations and other outcomes were registered but are not reported on here as analysis has yet to be finished. For consistency we have used here the hypothesis identifiers from the pre-registration.

(3) This process uses the python package *sexmachine*, which reports a 5-option likert value, from mostly female to mostly male, with androgynous as the balancing option. Only users with names which mapped to mostly male or mostly female were retained for analysis.

References

Bell, A. E., Spencer, S. J., Iserman, E., & Logel, C. E. (2003). Stereotype threat and women's performance in engineering. *Journal of Engineering Education*, 92(4), 307-312.

Brooks, C., & Gardner, J. (2017, November 30). Situational Gender Bias in Massive Open Online Courses. Retrieved from osf.io/7dxx2

Cheryan, S., Plaut, V. C., Davies, P. G., & Steele, C. M. (2009). Ambient belonging: How stereotypical cues impact gender participation in computer science. *Journal of personality and social psychology*, 97(6), 1045.

Cohen, J. (1994). The earth is round (p<. 05). *American psychologist,* 49(12), 997.

Gardner, J. & Brooks, C. (in press). Student Success Prediction in MOOCs. *User Modeling and User-Adapted Interaction.* doi: 10.1007/s11257-018-9203-z

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, 5(2), 189-211.

Kizilcec, R. and Brooks, C. (2017). Diverse Big Data and Randomized Field Experiments in Massive Open Online Courses. In Lang, C., Siemens, G., Wise, A. F., and Gaevic, D., editors, *The Handbook of Learning Analytics*, pages 211–222. Society for Learning Analytics Research (SoLAR), Alberta, Canada, 1 edition.

Kizilcec, R. F., & Cohen, G. L. (2017). Eight-minute self-regulation intervention raises educational attainment at scale in individualist but not collectivist cultures. *Proceedings of the National Academy of Sciences*, 201611898.

Kizilcec, R. F., Saltarelli, A. J., Reich, J., & Cohen, G. L. (2017). Closing global achievement gaps in MOOCs. *Science*, 355(6322), 251-252.

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

Kruschke, J. K. (2013). *Bayesian estimation supersedes the t test.* Journal of Experimental Psychology: General, 142(2), 573.

Kruschke, J. K., & Liddell, T. M. (2015). The Bayesian new statistics: two historical trends converge. *SSRN Electronic Journal.*

Lippa, R. A. (2010). Gender differences in personality and interests: When, where, and why? *Social and personality psychology compass*, 4(11), 1098-1110.

Rudman, L. A., & Goodwin, S. A. (2004). Gender differences in automatic in-group bias: Why do women like women more than men like men? *Journal of personality and social psychology*, 87(4), 494.

Stephenson, N. (1995). *The Diamond Ago: Or, A Young Lady's Illustrated Primer*. New York: Bantam Books.

Wasserstein, R. L., & Lazar, N.A. (2016). The ASA's Statement on P-Values: Context, Process, and Purpose. *The American Statistician* 70 (2). Taylor & Francis:129–33.

Acknowledgements

The authors would like to thank the anonymous reviewers for their insights and suggests on the paper, as well as all of the production staff and volunteer models from the Office of Academic Innovation at the University of Michigan. This project was funded in part by the Michigan Institute for Data Science (MIDAS) challenge award for Holistic Modeling of Education (HOME).