The Effects of Avatar Gender on Ratings of Reconstructed Teaching

Yilu Sun[1,3*], **Yejoon Yoo**[1], **Hyunju Kim**[1], **Yeonju Jang**[1], **Xianyi Li**[1], **Rimjhim Singh**[1], **Yuqing Wu**[1], **Wei Yang**[1], **Aleshia Hayes**[2], **and Andrea Stevenson Won**[1]

[1]Cornell University, Ithaca, NY 14853, USA
[2]University of North Texas, Denton, TX 76203, USA

ABSTRACT

Classroom simulators allow teachers to improve their teaching skills through practice and self-reflection. Teachers' nonverbal behavior can reveal their biases, set the tone of the classroom, and in these ways, influence student behavior positively or negatively. Virtual reality simulators allow teacher behavior to be tracked and transformed, creating a tool for improvement. However, in order to create tools to help teachers *improve* their nonverbal behavior, we must also understand how students *interpret* nonverbal behaviors, and to what extent "appropriate" teaching behaviors are determined by the identity of the teacher. In an online study using the reconstructed movements of participants with teaching experience performing a teaching task, we replaced the actual appearance of the teacher-participants with avatars of the same or different genders. We then identified differences in ratings of teaching quality based on whether the avatar matched or did not match the gender of the original performer of the actions.

Introduction

Nonverbal behavior is of key importance when teaching in the classroom[1]. While teachers use spoken language to communicate information, their non-verbal behaviors set the tone of the classroom and influence student behaviors[1]. Additionally, these non-verbal cues reveal unconscious biases that can potentially harm the classroom atmosphere[2].

Teacher training and self-reflection tools have aimed to improve nonverbal behavior because increased use of non-verbal communication has been associated with more effective education and enhanced academic progress for students[3]. Virtual Reality (VR) allows teachers to reflect on their nonverbal behaviors, especially as they relate to students, in unique ways[4]. Because their nonverbal behavior can be recorded and rendered in ways that make it more salient, teachers can use VR tools to reflect on their influential nonverbal behaviors such as how evenly they distribute their gaze, nonverbal synchrony between teachers and students, and proximity to students in different parts of the classroom[5,6]. Teachers can learn to efficiently deploy their nonverbal behaviors in a way that encourages a positive classroom experience for all users, and avoids showing bias towards one group of students or another[7].

However, biases in the classroom apply to both teachers and students. Teachers may be biased in rating student behavior[8], but using student ratings as a metric for teacher performance is also problematic, due to biases in *student* ratings[9]. When making recommendations for teachers' nonverbal behavior, one key assumption is that this advice would be broadly useful. However, is it accurate to say that there are universal standards for nonverbal behaviors of teachers? Or, are some behaviors viewed positively in male-presenting teachers but negatively in female-presenting teachers, as work on the gendered nature of nonverbal behavior might suggest? If so, then this discrepancy will inform both how we assess, and how we teach, nonverbal behavior during classroom teaching.

To answer these questions, we use the datasets of eight participants (four men and four women) who were videotaped while they were giving a short lesson to a small class of research assistants acting as students. From these videotapes, their teaching sessions were reconstructed in 3D. The datasets of the teacher-participant's gestures allowed us to replace the body of the teacher-participant who had originally performed the movements with avatar bodies of the same or different gender. We then created brief videos of these animations and presented them to a new set of participants for rating their nonverbal behavior.

Below, we briefly review literature on the importance of nonverbal behavior in teaching and teaching simulators. We discuss the well-known issues of biases in teaching evaluations, and how these complicate relying on student ratings as measures of "good" or "bad" teacher behavior. We briefly review differences in nonverbal behavior across genders. We then report the results of an online experiment examining how social factors, in this case, the gender of the avatar, affected how viewers judged the reconstructed animations of nonverbal behavior of teachers. We discuss how this can inform the use of virtual environments as teacher training tools.

In a virtual environment, users can observe and communicate with nonverbal behavior, use spatial awareness when interacting with objects[4], and present identity information[10] in ways similar to face-to-face settings. VR is utilized in a wide variety of educational scenarios as a crucial technology which can help students engage with abstract or unavailable materials, such as virtual monuments in the study of culture heritage[11], or surgical simulations in biomedical training[12]. These qualities have made virtual environments attractive as teaching tools, especially during the recent pandemic[13], and interest in virtual classrooms remains even today.

The simulation of teachers' behavior is also used to generate individual teaching reflections in addition to the study of teaching patterns in general. Richter and colleagues captured teachers' behavior in a virtual classroom, allowing the teacher to take the perspective of his or her own student. Their findings suggest this to be a more efficient way of reflecting on teaching than just watching a video of teaching[14]. Similarly, recorded movement patterns from a virtual classroom shed light on preservice teachers' behavior[15].

However, even when teachers are teaching in a physical classroom, virtual, augmented and mixed reality environments can still be useful tools, because they allow teacher behavior to be recorded and rendered in unique ways that make teacher behavior more salient[16]; for example, by allowing teachers to visualize their gaze[17]. New techniques of 3D reconstruction technology allow the capture and replaying of teachers' movements, including in virtual environments, which facilitate the study of teachers' behaviors and learning outcomes[18, 19]. These new techniques make the transfer of information about teaching behavior between physical and virtual classrooms is a useful area of research.

Nonverbal behavior is a key tool for effective teaching[20]. Effective nonverbal communication on the part of teachers is correlated with student success[21]. Ineffective communication can accidentally convey teacher's biases and emotional states[22]. Such ineffective communication harms individual students and can damage the classroom climate, during lessons, assessment, and classroom management. Students are aware of teacher bias, especially as expressed through nonverbal behaviors, and in turn, students' awareness of teachers' biases can harm learning[22], so much so that it has been proposed as an important component of teacher evaluation[23]. Nonverbal behavior can also provide information about biases of the communicator. For example, in a study by Richeson and Shelton, participants were able to infer actors' self-reported racial bias based on brief video recordings[24]. Students can also pick up this information from their own teachers' nonverbal behavior, with corresponding detrimental effects on learning[2]. Thus, we asked *RQ1. How will nonverbal behavior influence viewers' ratings of teacher effectiveness?*

Students' evaluation of teachers is also influenced by many external factors. Factors such as student motivation and expected grades, course level, academic field, etc., can bias student rating[25]. In addition, aspects of the teacher's identity can afefct ratings. Racial bias is present in teacher evaluations, and age and even physical appearance can affect students' evaluations of teachers[26, 27]. The influence of teacher gender[28] is particularly relevant to student ratings of teachers. Male students' biases may lead them to prefer male professors even when teachers demonstrate the same level of teaching effectiveness[29]. At the same time, gender stereotypes persist with male and female teachers. Although the gender bias associated with female teachers may lead them to be perceived as warmer and more powerful, they may then be evaluated by more rigorous standards than male teachers[30]. Male teachers may also thought to be more knowledgeable[31]. Therefore, when students are asked to rate teachers, male teachers often score higher. This gender bias not only exists in the physical classroom, but also in the online teaching environment. A recent study found that while evaluations of male teachers remained roughly the same in an online teaching semester, evaluation scores for female teachers declined, even after controlling for other factors such as poor performance of female teachers and students' classification of courses[32]. This leads to our second research question: *RQ2. Will animations with male avatars be rated differently than animations with female avatars, even when the avatar behavior (animation) is the same?*

Existing research has consistently highlighted gender disparities in both the expression and interpretation of emotions through nonverbal behavior. Earlier research has found that women tend to exhibit more expressive behavior, including increased smiling, and demonstrate higher levels of accuracy in both expressing and reading nonverbal cues compared to men[33]. These differences are influenced by underlying emotional states and societal gender norms. Identifiable differences in gender can even be found in such sparse stimuli as point-light displays, in which skeletal movement is conveyed through lights placed on the joints of an actor so that only a few points indicate the position of the body[34, 35]. Such characteristic differences in movement alone may reveal the gender of the original 'behaver' to participants. in turn, this may conflict with expectations arising from the perceived gender of the avatar. This leads us to our third research question: *RQ3. Are there differences by the gender of the original teacher whose behaviors are being reconstructed, in how participants rate reconstructed teaching behavior?*

Some work investigating gender-related bias in student evaluations of teachers changed the names of teachers in order to examine how gender cues might change ratings[36]. However, it is difficult to assess nonverbal behavior without revealing teacher gender. The ability to replace the physical appearance of the teacher with a different-looking avatar is one way VR can researchers investigate the effect of apparent social roles on observers. This technique has been used in other domains, often in a hospital setting. For example, Mast and colleagues conducted a study comparing people's responses to male vs.

female virtual physicians, examining whether patients' preferred communication styles of physicians changes based on whether the physician's avatar is male or female[37]. Other work has examined how varying the demographics of virtual patients can change decisions about pain management[38]. A key aspect of this type of manipulation is that the *behavior* of the avatar can be held constant, but the *appearance* of the avatar can be varied, allowing researchers to understand the effects of the social roles evoked by avatar appearance. For example, if a patient-avatar lists their symptoms, and complains of pain using the same animation and script, but participants are less likely to prescribe medications when the avatar appears female than when the avatar appears male, then this implies that bias (based on the avatar's appearance) is a factor. However, this raises an interesting question when the behavior animating a given avatar may in itself contain gendered cues; for example, when it is reconstructed from videos of actual human participants. This leads us to our fourth research question: *RQ4: Is there an interaction between avatar gender and the gender of the original teacher whose behaviors are being reconstructed, such that participants rate avatars which are congruent with the original actor's gender differently than avatars which do not match (are incongruent with) the original actor's gender?*

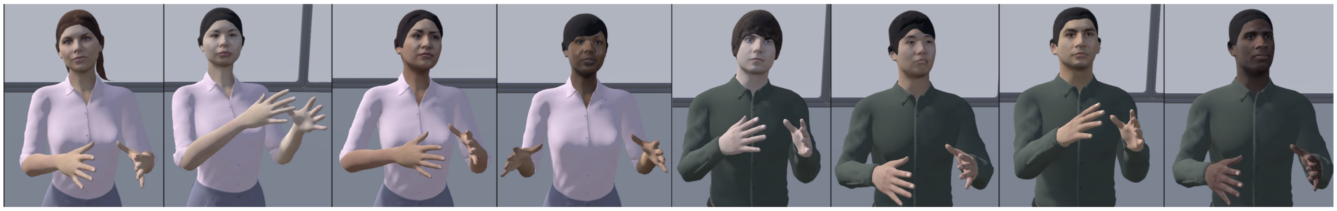

Figure 1. Eight avatars of different genders and ethnicities animated by recorded participant movements.

Results

All analyses were conducted in R[39].

RQ1. How will nonverbal behavior influence viewers' ratings of teacher effectiveness? After each video, we asked participants how different aspects of nonverbal behavior influenced their ratings. Given that people may also have different expectations about male and female behavior, we examined the interaction between avatar gender and different types of nonverbal behavior (head, hands, and body movement). We found that all three types of movement were related to participants' ratings on teaching behavior, but in different ways. The more influenced participants reported being by hand movements, the higher they rated the teaching behavior portrayed ($F(1, 718.90) = 27.843$, $p < 0.001$, 95% CI [0.112, 0.301]). However, there was no significant effect of avatar gender ($p > .05$). Similarly, the more participants reported being influenced by head movements ($F(1, 758.99) = 23.641$, $p < 0.001$, 95% CI [0.122, 0.276]), the higher they rated teacher behavior overall, and the interaction with avatar gender was also significant ($F(1, 636.23) = 3.947$, $p = 0.047$, 95% CI [-0.207, -0.001]), such that when participants were influenced by female avatars' head movements, they rated their teaching behavior slightly higher than male avatars. When examining body movements, there were also statistically significant relations between teacher behavior ratings and avatar gender ($F(1, 616.61) = 5.830$, $p = 0.016$, 95% CI [0.086, 0.827]), as well as a significant interaction of avatar gender and body movement influence ($F(1, 621.97) = 8.196$, $p = 0.004$, 95% CI [-0.262, -0.049]). Body movements did not change participants' ratings of male avatar's teaching behavior, but they did improve their rating of female avatars' teaching behavior.

RQ2. Are there differences by avatar gender in how participants rate reconstructed teaching behavior? Avatar did not influence ratings of the teaching behavior. ($F(1, 572) = 0.798$, $p = 0.372$, 95% CI [-0.174, 0.065]).

RQ3: Are there differences by the gender of the original teacher whose behaviors are being reconstructed, in how participants rate reconstructed teaching behavior? Male animations were rated slightly higher than the female animations ($F(1, 572) = 3.022$, $p = 0.083$, 95% CI [-0.014, 0.225]). However, given that the sample was small, this result should be considered with caution.

RQ4: Is there an interaction between avatar gender and the gender of the original teacher whose behaviors are being reconstructed, such that participants rate mismatched and congruent avatar-original teacher gender differently? We found a main effect of congruence on participants' rating of the teaching behavior ($F(1, 572) = 6.617$, $p = 0.010$, 95% CI [0.037, 0.275]) such that participants rated the teaching behavior higher when the avatar's gender was congruent with the teacher-participant whose movements created the animation. In order to better understand this effect, we ran a second model including the interaction of avatar gender and teacher-participant gender, with participant ID as a random effect. We found a statistically significant interaction between avatar and teacher gender ($F(1, 570) = 6.639$, $p = 0.010$, 95% CI [0.075, 0.549]). This effect was primarily driven by male avatars, who were rated significantly lower on teaching behavior when embodying the recorded movements of female teacher-participants.

Other Factors Affecting Teaching Behavior Ratings. The more participants felt they resembled the avatar, the better they rated the teaching behavior ($F(1, 737.77) = 29.508$, $p < 0.001$, 95% CI [0.086, 0.289]). Also, more natural-appearing avatars received higher teaching behavior scores ($F(1, 759.45) = 242.945$, $p < 0.001$, 95% CI [0.322, 0.459]). Avatar congruency also related to naturalness ratings, with a marginally statistically significant result such that congruent avatars were viewed as slightly more "natural" ($F(1, 572) = 2.729$, $p = 0.099$, 95% CI [-0.022, 0.258]).

Ranking Teachers. We found no significant main effect of animation gender, avatar gender, or gender congruence on how participants ranked the avatars (all p's $>$ .05). The correlation between ranking and the attractiveness ratings of the original photographs that were used to create the avatars was positive (r = .25) but not statistically significant.

Discussion

In this online study, we used recorded teaching behavior to create 64 short videos in which eight different male and female avatars performed each of eight animations. We found that when participants reported being influenced by different nonverbal behaviors, they rated the teaching effectiveness of the behaviors depicted in the videos higher. However, contrary to our expectations, the gender of the avatar mostly did not affect viewer's ratings of teaching effectiveness. However, we did find a statistically significant effect of *congruence*; participants rated animations as demonstrating less effective teaching behaviors when the avatar's gender was incongruent with the gender of the original teacher-participant; especially when a male avatar was animated by a female participants' movements. We also found that participants' perceptions of self-similarity with the avatar also increased their ratings of teaching effectiveness. This aligns with other findings on affinity bias, in which people exhibit a preference for individuals perceived to be similar. These relationships can be complex, as demonstrated in a recent study in which gender-affinity bias in which students evaluated teachers of their own gender better than those who were not of their gender[40].

Our findings extend previous work on biases in ratings of teaching behavior. They also have implications both for how researchers may make recommendations on teaching behavior, and on how avatar-swapping can be used to understand the effects of social roles on such ratings, although we wish to highlight that we are *not* recommending that the solution to rater bias is to adhere more strictly to perceived gender norms. While more research is needed, we suggest that when manipulating avatar appearance to examine bias, care must be taking to make sure that discrepancies between the avatar and the source of the avatar behaviors do not cause unexpected effects. In addition, our results provide additional support for not relying too heavily on student ratings to make recommendations for teacher behavior.

This study has a number of limitations, which we hope can be addressed in future work. First, we used only eight animations, from four men and four women. While all of these animations were derived from participants who had experience as teaching assistants, these animations were not selected as being "good" or "bad" but varied naturally in quality. Second, the reconstructed videos did not include audio or facial expressions and had some anomalies; for example, the avatar hands would sometimes appear to pass through each other. While overall, the ratings of "naturalness" were medium, and participants were specifically instructed to focus on nonverbal behavior some participants still commented that the unmoving facial expressions Next steps could include using existing datasets that would allow us to animate the avatar's facial expressions [21], as this is an important (and potentially gendered) aspect of nonverbal behavior in teaching, and modifying audio through pitch modulation to create masculine and feminine versions of the same speech. Third, while we were intentional in creating a set of avatars who spanned different demographic categories and were rated as approximately equally attractive, a broader sample of avatars would allow us to examine other factors linked with bias, such as age, race and body type. Fourth, we did not examine participants' expectations of appropriate behavior for male and female teacherswhich could help us better understand our findings.

Conclusions

Our findings highlight potential challenges in using avatar-swapping techniques to isolate the effects of avatars from behaviors, as viewers may be sensitive to perceived incongruencies. However, by manipulating avatar embodiment to understand how avatar behavior intersects with we propose that future work can not only highlight biases, but point to a way to identify successful behaviors that may be more universal.

Methods

We conducted a within-participants study in which participants from a crowdsourced platform (Prolific[41]) each rated four silent videos of an avatar conducting a lecture for teaching effectiveness. Each video showed an avatar randomly chosen from a set of eight avatars created for this study. The avatar was performing the reconstructed nonverbal behavior of teacher-participants from a previous study who were videotaped delivering a brief lecture to a small class.

Participants

Participants (N=194) were recruited through Prolific. Three participants were excluded due to survey design errors. Among the 191 participants, 94 participants identified as female, 92 participants as male. Participants could select from more than one category for race/ethnicity, so the numbers per category below do not sum to the number of participants. Fifty-six participants selected Asian; 55 participants, Black or African American; 40 participants Hispanic or Latino; four participants, Native American or Alaska Native, and 45 participants, White. One person selected the category "Other." Seventeen participants had graduated high school, 44 had at least some college education, 20 participants had a 2-year degree, 80 had a 4-year degree, 10 had a professional degree, 18 had a master's degree, and two had a doctorate. When asked "When was the last time you observed a teacher in front of the classroom like the videos?" 12 participants said they had never seen this, 66 participants said within one year, 21 participants said 1-2 years, 40 said 3-5 years, 23 said 6-10 years, 28 said more than 10 years, and one selected other, writing "I see this daily". Ten participants were between 18-20 years of age, 77 between 21-30, 52 between 31-40, 31 between 41-50, and 20 participants were 51 or above. One participant reported being aged "1" which we interpreted as a typographical error. Seventy participants did not have previous experience with video or computer games with avatars, while 121 participants had. All procedures were approved by the Institutional Review Board, and all participants signed informed consent.

Materials

Avatar Creation

To create the avatars that would be animated using the reconstructed motions of the teacher-participants, we aimed to have a diverse set of avatars which participants would interpret as female or male. To do so, we created a set of avatars using head shots of four males and four females from the Chicago Face Database (originally described here and later expanded[42,43]. This database consists of hundreds of photographs of adults who self-identified as belonging to four racial/ethnic categories (Asian, Black, Hispanic/Latino, or White). Norming data for each of these pictures was derived by having participants on a number of measures, including how typically male or female the faces appear, how typical they appear of their self-identified racial/ethnic group, and how attractive they appear. In order to create a diverse group of avatars that would be viewed as gender-representative, we first created a sub-selection of photographs which were rated as appearing "typically female" or "typically male." We then further divided this selection into avatars that were rated as "typical" for their identified racial/ethnic group. Finally, we selected the photographs within each subset of race/ethnicity and gender which were closest to the average attractiveness rating of 3.5 on a scale of 1-7. The head shots were then input into Avatar SDK (https://avatarsdk.com) to create the full body, along with clothing. Then the full body avatars were exported as FBX files and imported into Unity to be animated with the movement data collected from teacher recordings in the section below. Figure 1 shows each avatar.

Animations of Teachers

The experimental stimuli of the teacher animation movement data were derived from a prior study in which 29 participants (all graduate students with teaching experience) were videotaped while conducting a brief lecture to a small group of research assistants. These videos were then reconstructed to create 3D skeletal data[44].Each of the eight avatars was then applied to each dataset to create eight animations for each dataset. The animations were derived from the first minute of their reconstructed lectures, following Ambady and Rosenthal's work finding consistency in ratings of teacher nonverbal behavior from less than a minute of videotaping[45]. In other words, we created eight one-minute animations of each teacher-participant's movement data file, for each of the eight avatars we generated. In congruent animations, the gender of the avatar and the original participant matched; for example, if a female avatar performed a female teacher-participant's reconstructed movements. In incongruent conditions, they differed; for example, if the same animation was performed by a male avatar. This created four video conditions: *Female Animation with Congruent Avatar, Female Animation with Incongruent Avatar, Male Animation with Congruent Avatar,* and *Male Animation with Incongruent Avatar*. These reconstructions did not include facial animation or audio, in order to reduce confounds from the content and audio qualities of the teacher-participants' speech.

Video Rating and Selection

Once all the videos were created and animated using avatars, five of the authors rated the videos for artifacts that might make the videos appear uncanny. The movement files of the 4 male and 4 female teaching assistants that were rated as least uncanny were selected to be in the final survey questions. The authors evaluated each video by assigning a rating on a scale of 1 to 7, where 1 indicated a strong agreement with the statement "The instructor's avatar is uncanny", and 7 indicated strong disagreement. The eight least uncanny videos were selected. Since every teaching assistants' movement is embodied by eight avatars, we have 32 videos created from the movement of the four male participants and 32 videos created from the movements of the four female participants.

Experiment Flow

Participants were recruited from the online platform Prolific[41]. We conducted a pilot study of 48 participants, using open enrollment approach that permitted all participants to take part in the survey. This resulted in a majority of White participants. From these data, we conducted a power analysis using G*Power3[46] which suggested a sample of 151 participants. In order to obtain a balanced sample, we used the Prolific 'simplified ethnicity' screener settings "Asian", "Black", "Mixed", "Other" and "White". We recruited in four waves using the"Black", "Asian", and "Other" settings and one open enrollment, resulting in 192 participants. Thus, each video version was viewed by approximately six people.

After informed consent, participants read: "Below is a one-minute video showing the recorded behavior of a teacher presenting a lecture to a small class. These are real recorded movements, but we have used a randomly selected avatar to represent the teacher to preserve their anonymity. Please note, this video has no sound, and no facial animation. We ask you to focus on the nonverbal behavior (gestures, posture, etc.) of the teacher only. After you have watched the video, we will ask you some questions about your impressions of their teaching based on their nonverbal behavior."

Participants were then presented with four videos, in random order, representing the four experimental conditions described above. No participant saw the same avatar or animation twice; i.e., each animation was derived from a different video source and no avatars were reused for each participant. After each video, participants completed a brief questionnaire. After they had watched all four videos and completed all four questionnaires, they were shown pictures of each of the avatars they saw and were asked to rank them on their nonverbal behavior.

Measures

Ratings of teaching behavior

Participants were asked to rate on a 7-point scale (1 = strongly disagree, 7 = strongly agree) on the following items from[47]:

1. The instructor created a climate of mutual trust and respect in classroom.
2. The instructor maintained a classroom setting that minimizes disruption.
3. The instructor created a friendly and supportive classroom environment.
4. The instructor ensured students' participation in the learning process.
5. The instructor encouraged students to interact respectfully.

The results of these five questions were averaged to create the teaching behavior metric (M = 4.745, SD = 1.044).

What influences participants ratings?

Participants were asked to rate on a 5-point scale on how much the avatar's hand gestures (M = 3.683, SD = 1.055), head movements (M = 2.893, SD = 1.218), body movements (M = 3.295, SD = 1.176), and appearance (M = 2.054, SD = 1.181) influenced their ratings respectively, with 1 = None at all, 5 = A great deal. Participants were also asked "How much do you think the avatar resembles you?" on a 5-point scale (1 = None at all, 5 = It looks exactly like me) (M = 1.681, SD = 0.954). Finally, they were asked "Did the avatar's behavior appear natural or unnatural?" (1 = Very unnatural, 5 = Very natural) (M = 3.122, SD = 1.261).

Ranking Teachers

After the participants rated all four videos, they were presented with the images of the four avatars whom they had seen in videos, and were asked "Please rate the four teachers you saw from worst to best with 1 being worst and 4 being best. Please note, the pictures may not appear in the order that you saw the animations!"

References

1. Babad, E. Teaching and Nonverbal Behavior in the Classroom. 817–827, DOI: 10.1007/978-0-387-73317-3_52 (2009).

2. Babad, E., Bernieri, F. & Rosenthal, R. Nonverbal communication and leakage in the behavior of biased and unbiased teachers. *J. Pers. Soc. Psychol.* **56**, 89–94, DOI: 10.1037/0022-3514.56.1.89 (1989). Place: US Publisher: American Psychological Association.

3. BAMBAEEROO, F. & SHOKRPOUR, N. The impact of the teachers' non-verbal communication on success in teaching. *J. Adv. Med. Educ. & Prof.* **5**, 51–59 (2017).

4. Hindmarsh, J., Fraser, M., Heath, C., Benford, S. & Greenhalgh, C. Object-focused interaction in collaborative virtual environments. *ACM Transactions on Comput. Interact.* **7**, 477–509, DOI: 10.1145/365058.365088 (2000).

5. Shaikh, O., Sun, Y. & Stevenson Won, A. Movement Visualizer for Networked Virtual Reality Platforms. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 681–682, DOI: 10.1109/VR.2018.8446398 (2018).

6. Kale, U. Levels of interaction and proximity: Content analysis of video-based classroom cases. *The Internet High. Educ.* **11**, 119–128, DOI: 10.1016/j.iheduc.2008.06.004 (2008).

7. Hayes, A. T., Hardin, S. E. & Hughes, C. E. Perceived Presence's Role on Learning Outcomes in a Mixed Reality Classroom of Simulated Students. In Shumaker, R. (ed.) *Virtual, Augmented and Mixed Reality. Systems and Applications*, Lecture Notes in Computer Science, 142–151, DOI: 10.1007/978-3-642-39420-1_16 (Springer, Berlin, Heidelberg, 2013).
8. Mason, B. A., Gunersel, A. B. & Ney, E. A. Cultural and ethnic bias in teacher ratings of behavior: a criterion-focused review. *Psychol. Sch.* **51**, 1017–1030 (2014).
9. Mitchell, K. M. W. & Martin, J. Gender Bias in Student Evaluations. *PS: Polit. Sci. & Polit.* **51**, 648–652, DOI: 10.1017/S104909651800001X (2018). Publisher: Cambridge University Press.
10. Ratan, R. & Hasler, B. Exploring Self-Presence in Collaborative Virtual Teams. *PsychNology J.* **8**, 11–31 (2010).
11. Goncalves, N. Educational use of 3d virtual environments: primary teachers visiting a romanesque castle. *Recent research developments learning technologies* 427–4331 (2005).
12. Juanes, J. A., Ruisoto, P. & Barros, P. Technological innovations in biomedical training and practice. In *Proceedings of the Sixth International Conference on Technological Ecosystems for Enhancing Multiculturality*, 348–350 (2018).
13. Won, A. S., Bailey, J. O. & Yi, S. Work-in-progress—learning about virtual worlds in virtual worlds: How remote learning in a pandemic can inform future teaching. In *2020 6th International Conference of the Immersive Learning Research Network (iLRN)*, 377–380 (IEEE, 2020).
14. Richter, E., Hußner, I., Huang, Y., Richter, D. & Lazarides, R. Video-based reflection in teacher education: Comparing virtual reality and real classroom videos. *Comput. & Educ.* **190**, 104601 (2022).
15. Huang, Y., Richter, E., Kleickmann, T., Scheiter, K. & Richter, D. Body in motion, attention in focus: A virtual reality study on teachers' movement patterns and noticing. *Comput. & Educ.* 104912 (2023).
16. Hayes, A. T., Hughes, C. E. & Bailenson, J. Identifying and Coding Behavioral Indicators of Social Presence With a Social Presence Behavioral Coding System. *Front. Virtual Real.* **3** (2022).
17. Bailenson, J. N. *et al.* The use of immersive virtual reality in the learning sciences: Digital transformations of teachers, students, and social context. *The J. Learn. Sci.* **17**, 102–141 (2008).
18. Ahuja, K. *et al.* Edusense: Practical classroom sensing at scale. *Proc. ACM on Interactive, Mobile, Wearable Ubiquitous Technol.* **3**, 1–26 (2019).
19. Ahuja, K. *et al.* Classroom digital twins with instrumentation-free gaze tracking. In *Proceedings of the 2021 chi conference on human factors in computing systems*, 1–9 (2021).
20. Altun, M. An underestimated tool: Body language in classroom during teaching and learning. *Int. J. Soc. Sci. & Educ. Stud.* **6**, 155–170 (2019).
21. Bambaeeroo, F. & Shokrpour, N. The impact of the teachers' non-verbal communication on success in teaching. *J. advances medical education & professionalism* **5**, 51 (2017).
22. Babad, E., Bernieri, F. & Rosenthal, R. Nonverbal communication and leakage in the behavior of biased and unbiased teachers. *J. personality social psychology* **56**, 89 (1989).
23. Babad, E., Sahar-Inbar, L., Hammer, R., Turgeman-Lupo, K. & Nessis, S. Student Evaluations Fast and Slow: It's Time to Integrate Teachers' Nonverbal Behavior in Evaluations of Teaching Effectiveness. *J. Nonverbal Behav.* **45**, DOI: 10.1007/s10919-021-00364-4 (2021).
24. Richeson, J. A. & Shelton, J. N. Brief Report: Thin Slices of Racial Bias. *J. Nonverbal Behav.* **29**, 75–86, DOI: 10.1007/s10919-004-0890-2 (2005).
25. Cashin, W. E. Student ratings of teaching: A summary of the research. idea paper no. 20. (1988).
26. Arbuckle, J. & Williams, B. D. Students' perceptions of expressiveness: Age and gender effects on teacher evaluations. *Sex Roles* **49**, 507–516 (2003).
27. Goebel, B. L. & Cashen, V. M. Age, sex, and attractiveness as factors in student ratings of teachers: A developmental study. *J. Educ. Psychol.* **71**, 646 (1979).
28. Laube, H., Massoni, K., Sprague, J. & Ferber, A. L. The impact of gender on the evaluation of teaching: What we know and what we can do. *Nwsa J.* 87–104 (2007).
29. Boring, A. Gender biases in student evaluations of teaching. *J. public economics* **145**, 27–41 (2017).
30. Bennett, S. K. Student perceptions of and expectations for male and female instructors: Evidence relating to the question of gender bias in teaching evaluation. *J. educational psychology* **74**, 170 (1982).

31. Miller, J. & Chamberlin, M. Women are teachers, men are professors: A study of student perceptions. *Teach. Sociol.* 283–298 (2000).

32. Ayllón, S. Online teaching and gender bias. *Econ. Educ. Rev.* **89**, 102280 (2022).

33. Hall, J. A., Carter, J. D. & Horgan, T. G. Gender differences in nonverbal communication of emotion. *Gend. emotion: Soc. psychological perspectives* 97–117 (2000).

34. Barclay, C. D., Cutting, J. E. & Kozlowski, L. T. Temporal and spatial factors in gait perception that influence gender recognition. *Percept. & psychophysics* **23**, 145–152 (1978).

35. Pollick, F. E., Kay, J. W., Heim, K. & Stringer, R. Gender recognition from point-light walkers. *J. Exp. Psychol. Hum. Percept. Perform.* **31**, 1247 (2005).

36. Macnell, L., Driscoll, A. & Hunt, A. What's in a Name: Exposing Gender Bias in Student Ratings of Teaching. *Innov. High. Educ.* DOI: 10.1007/s10755-014-9313-4 (2014).

37. Schmid Mast, M., Hall, J. A. & Roter, D. L. Disentangling physician sex and physician communication style: Their effects on patient satisfaction in a virtual medical visit. *Patient Educ. Couns.* **68**, 16–22, DOI: https://doi.org/10.1016/j.pec.2007.03.020 (2007).

38. Wandner, L. D. *et al.* The impact of patients' gender, race, and age on health care professionals' pain management decisions: An online survey using virtual human technology. *Int. J. Nurs. Stud.* **51**, 726–733, DOI: https://doi.org/10.1016/j.ijnurstu.2013.09.011 (2014).

39. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2021).

40. Binderkrantz, A. S. & Bisgaard, M. A gender affinity effect: the role of gender in teaching evaluations at a Danish university. *High. Educ.* DOI: 10.1007/s10734-023-01025-9 (2023).

41. Prolific. What is Prolific? (2023).

42. Ma, D. S., Correll, J. & Wittenbrink, B. The chicago face database: A free stimulus set of faces and norming data. *Behav. research methods* **47**, 1122–1135 (2015).

43. Ma, D. S., Kantner, J. & Wittenbrink, B. Chicago face database: Multiracial expansion. *Behav. Res. Methods* **53**, 1289–1300 (2021).

44. Easymocap - make human motion capture easier. Github (2021).

45. Ambady, N. & Rosenthal, R. Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *J. personality social psychology* **64**, 431 (1993).

46. Faul, F., Erdfelder, E., Buchner, A. & Lang, A.-G. Statistical power analyses using g* power 3.1: Tests for correlation and regression analyses. *Behav. research methods* **41**, 1149–1160 (2009).

47. Akram, M. & Zepeda, S. J. Development and validation of a teacher self-assessment instrument. *J. Res. & Reflections Educ. (JRRE)* **9** (2015).

Acknowledgements (not compulsory)

Acknowledgements should be brief, and should not include thanks to anonymous referees and editors, or effusive comments. Grant or contribution numbers may be acknowledged.

Author contributions statement

Must include all authors, identified by initials, for example: A.A. conceived the experiment(s), A.A. and B.A. conducted the experiment(s), C.A. and D.A. analysed the results. All authors reviewed the manuscript.

Additional information

To include, in this order: **Accession codes** (where applicable); **Competing interests** (mandatory statement).

The corresponding author is responsible for submitting a competing interests statement on behalf of all authors of the paper. This statement must be included in the submitted article file.