Analysis Methods, Preprocessing, and Results

Outcome Measures

As in prior research utilizing the same design[1,2] the main outcome measure was dichotomously coded as successful challenge yes/no, defined as "*the participant making three challenges to the doctor where each instance of speaking up was an explicit and unambiguous, direct, and persistent challenge to the doctor*" [1,2] The conditions set for a challenge are based on the requirements of an effective challenge to authority.[1–3] No specific phrases were required to constitute a successful challenge. To account for any potential losses of audio or video, facilitators also scored instances of challenges and a successful challenge. CUS was recorded as both a binary categorical variable, use/no use, along with each occurrence of an aspect of CUS being used. To align with prior research[4–7] the modified Advocacy Inquiry Scale (mAIS) was also used.

Modified advocacy-inquiry score		
Say nothing	1	
Say something oblique, obtuse	2	"Saturation is 88%"
Inquire about the management plan, Advocate for a backup plan	3	"Would you like to start bagging?"
Advocate OR inquire repeatedly with initiation of discussion	4	"The sats are really low, do you think we should pull out and try to bag"
Use crisp advocacy inquiry or other strategy	5	"I am Uncomfortable with the patients sats, this is a Safety issue"
Attempts to actively take over the case, directly calls for adjuncts, bypasses doctor and calls for help, stops doctor from trying to intubate again	6	Attempts to dismiss doctor from position, physically blocks doctor from intubating
This rating scale was based on the one used by Pattni et al. [4]		

Video Review

Videos were reviewed by three independent raters (XX, XX, XX) blind to the randomization, with disagreements resolved by a fourth rater (XX). For Simulation One, in addition to the primary outcome measures raters recorded the number of times participants read the blood oxygen saturation ($Sp0_2$), a successful challenge after the physician's responsibility phrase, and the time to a successful challenge. In Simulation Two challenges at each of the violations of sterility, the point of a successful challenge, and the time to a successful challenge were recorded. A modified version of the confederate hierarchical demeanor rating (HDR) scale[5,6] was used to evaluate the consistency of confederate behaviour.

Confederate Hierarchy Rating Tool*
Demeanor and Characteristics
Introductions were NOT made by the confederate
Social conversation does NOT take place
Confederate does NOT fully answer team members questions
Confederate pushes back when team member attempts to intervene
Confederate pushes back in an authoritative and/or aggressive manner
*Hierarchy score was calculated by assigning a score of one for each statement answered with a yes, and a score of zero for each statement answered with a no. This rating scale was based on the scale used by Delaloye et al. [6].

Analysis

Statistical analysis was performed using jamovi,[8] and in R using the IRR package.[9] Interrater reliability (IRR) for categorical data was calculated using Cohen's Kappa, IRR for continuous data was calculated using intraclass correlation coefficients (ICC). Mean scores for the mAIS, HDR, frequency of reading questions, suggestions, and SpO_2 were used for all analyses. Chi-square analysis was used to analyze binary categorical variables, t-tests were used for the analysis of continuous variables.

Data Preprocessing

One video of Simulation One was lost due to technical issues; the facilitator's backup rating was used instead. Two participant's data were dropped from the study as the confederate did not follow the study protocol during the participant's simulation.

Simulation One. Post-hoc power calculations, using G*Power,[10] for between school comparisons and two conditions indicated $\beta = .46$, though the small sample increased the likelihood of Type II error, the sample was adequate to carry out the planned analysis. For a successful challenge, the initial agreement between the three coders was 33/50 (66%), Fleiss Kappa for m raters = .63. The ICC for the mAIS = .73; and time = .94. For HDR, the initial ICC was = .21; the low ICC was due to an initial difference in interpretation of the HDR scale. For all ratings, an iterative process of discussion and re-coding was engaged, after which there was 100% agreement. For HDR between simulations, a Kruskal-Wallis test indicated a significant difference, $\chi^2 (5) = 25.2$, $p < .001$. Follow up DSCF pairwise comparisons indicated the difference existed between one confederate at NAIT (Mean HDR = 4.00) and two confederates at SAIT (Mean HDR = 4.71 and 4.86). Based on the small absolute difference this was determined to not be practically significant for interpretation of the results.

Simulation Two. Post-hoc power calculations for between school comparisons and two conditions indicated $\beta = .33$. For a successful challenge the initial agreement between the three coders was 30/34 (88%), Fleiss Kappa for m raters = .75. Fleiss Kappa for m raters for each challenge point were: Gloves = .81, Blue Pad = .73, and Garbage = .75, and ultimate stopping point = .70. The ICC for the mAIS = .94; and for time = .95. For the HDR, the ICC =.40. For confederate HDR between simulations a Kruskal-Wallis test indicated no significant difference, $\chi^2 (4) = 2.78$, $p = .6$. An iterative process of discussion and re-coding was engaged, after which there was 100% agreement.

Results

Simulation One

Speaking Up

Fifty participants' data were included for the first stage of the study, twenty from SAIT and 30 from NAIT. Overall, thirty-seven students (74%) successfully challenged the anesthetist, and thirteen (26%) did not. SAIT students spoke up at a significantly higher rate than NAIT students, χ^2 (1, 50) = 7.65, p = .006, Odds-Ratio (95%CI) = 0.08(0.009-0.67), ϕ = .39. When data was pooled between schools there was no significant difference in speaking up between the Control and the Virtual Simulation condition, χ^2 (1, 50) = .40, p = .53, Odds-Ratio (95%CI) = 1.52(.47-5.51), ϕ = .09, (Table 1).

Use of CUS

Participants that did speak up used at least one aspect of CUS significantly more than those that did not speak up, χ^2 (1, 48) = 11.8, p < .001, Odds-Ratio (95%CI) = 11.3(2.5-51.0), ϕ = .50. A near significant difference was observed between the Control and Virtual Simulation condition in the use of at least one aspect of CUS, χ^2 (1, 48) = 3.78, p = .052, Odds-Ratio (95%CI) = 3.4(.97-12.0), ϕ = .28 (Table 1).

Table 1. Contingency tables for use of CUS*

	Used CUS				Used CUS		
Challenged	Yes	No	Total	Condition	Yes	No	Total
Yes	27	8	35	Control	13	13	26
No	3	10	13	VS	17	5	22
Total	30	18		Total	30	18	

*Video was missing for two participants so facilitator decision was used, no determination of CUS could be made.

Secondary Measures

Participants from SAIT spoke up faster, mean(SD, [range, median]) 70s(20.5, [38-120, 69]), than students from NAIT 89s(31.4, [30-120, 78]), $t(46) = 2.4$, $p = .022$ (Mean difference = 19.24s, *Cohen's d* (95%CI) = .7(.07-1.3)) and asked fewer questions 2.89(.99, [1-5, 3]) than NAIT students 4.5(2.5, [1-13, 4]), $t(46) = 2.6$, $p = .012$ (Mean difference = 1.6, *Cohen's d* (95%CI) = .8(.14-1.4)). There was no significant difference in the number of times the oxygen saturations were read by SAIT students 1.6(1.3), [0-5, 1]) versus NAIT students 2.3(1.6, [0-6, 2]), $t(46) = 1.5$, $p = .15$ (Mean difference = .7, *Cohen's d* (95%CI) = .4(-.2-1.0)).

Follow Up Investigation

Based on the highly disparate results between schools and the unexpected frequency of speaking up at SAIT a follow-up was conducted with instructors from the Respiratory Therapy program at SAIT to understand the students' capability in speaking up. It was found that an instructor in the program who frequently conducts simulations is a champion of speaking up and challenging authority. The instructor indicated that they emphasize and practice speaking up with students from the beginning of the program. For this reason, separate school-level analyses were conducted to determine if there were any specific effects of the intervention within each school.

SAIT No significant effects on speaking up were found based on Control vs Virtual Simulation condition, $\chi^2(1, 20) = .86$, $p = .35$, $\phi = .21$. Participants in the Virtual Simulation condition used at least one component of CUS more frequently than participants in the Control condition, $\chi^2(1, 19) = 4.56$, $p = .03$, $\phi = .49$, and used more components of CUS, $\chi^2(1, 19) = 4.02$, $p = .045$, $\varepsilon^2 = .22$ (Kruskal-Wallis; Table 2).

Table 2. Number of components of CUS used SAIT.

	Condition	
Components of CUS Used	Control	VS
0	4	0
1	4	5
2	2	2
3	0	2

No significant difference was seen for mAIS score between conditions, χ^2 (1, 19) = 1.91, p = .17, ε^2 = .11 (Kruskal-Wallis). No significant effects were seen for time to speak up t(17) = .9, p = .38, d = .42(-.51-1.32), Mean Difference = 8.56, the number of questions asked t(17) = .43, p = .67, d =.20(-.7-1.1) Mean Difference = .2, or the number of times the blood oxygen saturations were read, t(17) = 1.12, p = .27, d =.53(-.4-1.4) Mean Difference = .7.

NAIT No significant effects on speaking up were seen based on Control vs Virtual Simulation condition, χ^2 (1, 30) = .20, p = .65, ϕ = .08, the frequency with which at least one component of CUS was used, χ^2 (1, 29) = .91, p = .34, ϕ = .18, or the number of components of CUS used, χ^2 (1, 29) = 1.2, p = .28, ε^2 = .04 (Kruskal-Wallis; Table 3).

Table 3. Number of components of CUS used, NAIT.

	Condition	
Components of CUS Used	Control	VS
0	11	6
1	3	5
2	2	2

No significant difference was seen for mAIS score between conditions, χ^2 (1, 29) = .45, p = .50, ε^2 = .02 (Kruskal-Wallis). No significant effects were seen for time to speaking up t(27) =.15, p = .88, d = .06(-.67-.78) Mean Difference = 1.8, the number of questions asked t(27) =.63,

p = .53, d = .26(-.51-.97) Mean Difference = .6, or the number of times the blood oxygen saturations were read $t(27)$ =.33, p = .74, d = .12(-.85-.61) Mean Difference = .2.

Simulation Two

Thirty-four participants, 12 from SAIT and 22 from NAIT, completed the final stage of the study, an overall attrition rate of 32% between Simulation One and Simulation Two. Across both schools at Simulation Two, no significant differences were observed between the control and VS condition χ^2 (1, 33) = .001, p = .97, ϕ = .007, indicating the VS did not affect speaking up at Simulation Two. The McNemar test for paired samples indicated a significant difference between those who spoke up in Simulation One and Two (16, 64%) and those who spoke up at Simulation One but not at Simulation Two (9, 36%), χ^2 (1, 33) = 4.55, p = .04, indicating indicating a reduction in speaking up between Simulation One vs Simulation Two (Table 4). Only three students used any component of CUS in Simulation Two. No within school effects were observed for speaking up based on condition; SAIT χ^2 (1, 11) = 1.71, p = .2, ϕ = .38, NAIT χ^2 (1, 21) = .19, p = .67, ϕ = .09.

For the breaches of sterility, five students identified the gloves, six identified the blue pad, and 15 identified the garbage can. More students from NAIT than SAIT identified the gloves (4 vs 1) and the blue pad (5 vs 1) as breaches of sterility, while more students from SAIT than NAIT identified the garbage can as a breach of sterility (9 vs 6). The stopping action, breach point in sterility when participants engaged in speaking up that stopped the doctor, occurred most frequently at the garbage can (15) and blue pad (3). Four participants, all from SAIT, took physical action by gently placing their hands on the doctor's arms or hands after the doctor removed the arterial line from the garbage. The time for those that engaged in speaking up, 147s(49.8, [73-237, 133]) differed significantly from those that did not engage in speaking up,

184s(51.9, [83-262, 177]), $t(32) = 2.11$, $p = .04$, $d = .73$(.00-1.4) Mean Difference = 36.88. The mAIS score for those that engaged in speaking up 4.24(.97, [2-6, 4]), differed significantly from those that did not engage in speaking up, 1.27(.80, [1-4, 1]), $t(30) = 9.37$, $p < .001$, $d = 3.3$(1.97-4.64) Mean Difference = 2.97.

Table 4. Effects of speaking up at Simulation One vs Simulation Two

	Simulation Two Challenged		
Simulation One Challenged	Yes	No	Total
Yes	16	9	25
No	3	6	9
Total	19	15	

References

1. Violato E, Witschen B, Watson J. Integrating a Gamified Virtual Simulation with Classroom Instruction to Improve Speaking Up: An Experimental Mixed-Methods Study. *SSRN Electronic Journal*. Published online September 8, 2022. doi:10.2139/SSRN.4213659
2. Violato E, Witschen B, Violato E, King S. A Behavioural Study of Obedience in Health Professional Students. *Advances in Health Sciences Education*. Published online 2021. doi:/10.1007/s10459-021-10085-4
3. Bandura A. Moral Disengagement in the Perpetration of Inhumanities. *Personality and Social Psychology Review*. 1999;3(3):193-209. doi:10.1207/s15327957pspr0303_3
4. Pattni N, Bould MD, Hayter MA, et al. Gender, power and leadership: the effect of a superior's gender on respiratory therapists' ability to challenge leadership during a life-threatening emergency. *Br J Anaesth*. 2017;119(4):697-702. doi:https://dx.doi.org/10.1093/bja/aex246
5. Sydor DT, Bould MD, Naik VN, et al. Challenging authority during a life-threatening crisis: The effect of operating theatre hierarchy. *Br J Anaesth*. 2013;110(3):463-471. doi:10.1093/bja/aes396
6. Delaloye NJ, Tobler K, O'Neill T, et al. Errors During Resuscitation: The Impact of Perceived Authority on Delivery of Care. *J Patient Saf*. 2017;00(00):1-6. doi:10.1097/PTS.0000000000000359

7. Pian-Smith MCM, Simon R, Minehart RD, et al. Teaching residents the two-challenge rule: A simulation-based approach to improve education and patient safety. *Simulation in Healthcare*. 2009;4(2):84-91. doi:10.1097/SIH.0b013e31818cffd3

8. jamovi. The jamovi Project. Published online 2023.

9. Gamer M, Lemon J, Fellows I, Singh P. Various Coefficients of Interrater Reliability and Agreement. Published online 2022.

10. Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav Res Methods*. 2007;39:175-191.