# Quantifying the benefits to students of faking your fluency in English in recorded lectures

Karol P. Binkowski and Greg Baker
Macquarie University, Australia
karol.binkowski@mq.edu.au

*Heavily accented speech can hinder learning, particularly for non-English speaking students who struggle with comprehension and engagement. This study explores whether AI-generated voice-modded lectures, mimicking native English speech, improve academic outcomes and engagement in lectures by non-native English-speaking instructors. The study was conducted in an Introductory Statistics large enrolment service unit and compared student performance across original and AI-enhanced lectures. Surveys were used to collect engagement feedback from both domestic and international students.*
*No significant performance difference was observed for international students. However, domestic students showed improved marks when exposed to synthetic lectures, highlighting this intervention's low-cost, scalable impact. This research offers practical insights into using AI voice-modding to improve engagement and equity in large, diverse classes.*

## INTRODUCTION

Previous research has shown that heavily accented speech can negatively impact student learning outcomes, particularly for those from non-English speaking backgrounds (McClure & Chen, 2024). Voice-modding technology, which uses AI-based speech synthesis and voice conversion techniques to modify the accent and prosody of speech, has shown promise in improving the intelligibility and accessibility of educational content (Zhao et al., 2021). However, the effectiveness of this technology in real-world educational settings has not been extensively studied. Our study complements the evolving paradigms in statistics education outlined by Ben-Zvi (2016), particularly in how technological interventions, such as AI-enhanced voice synthesis, can support statistical reasoning by reducing cognitive barriers to comprehension. This area remains under-researched due to ethical concerns about authenticity, consent, and potential bias against accents. Practical challenges in implementation and cultural sensitivity may also have limited exploration. As this intervention was trialled in a statistics unit, it offers a unique opportunity to explore how language clarity influences comprehension in a field that already challenges students with technical vocabulary, abstract concepts and reasoning.

Quantitative fields have a disproportionately high number of non-English-speaking lecturers, for example, in the U.S. (Sabharwal, 2011), with similar trends observed in developed countries, such as Australia. International enrolment bias, statistics, data science, and related quantitative disciplines attract many international students, especially from countries where English is not the primary language, including China, India, Iran, or Bangladesh. Many of these students go on to pursue academic careers. In STEAM fields, hiring is becoming increasingly international because skill sets like mathematics, statistics, and programming are easily transferable worldwide, despite language barriers.

### *Teaching context*

The Faculty of Science and Engineering offers a generalist degree, *Bachelor of Science*, with several majors run by departments and schools across the university. With an annual enrolment of approximately 1200 students in each of the two sessions, the Introductory Statistics unit comprises students enrolled in various predominantly non-statistical, STEAM and some non-STEAM disciplines. Designed for on-campus delivery with over three-quarters of students enrolled in in-person mode, the unit has been modified to accommodate online classes and hybrid live lectures. Binkowski (2023) provides more background on the unit, including the educational philosophy of Mastery Learning that underpins the assessment structure and analyses students' performance across various degree disciplines. In addition, they analysed the relationship between students' assessment performance, their time spent watching pre-recorded content, and their effort on practice quizzes. The unit is considered to have a high online student engagement rate, as Thurn (2023) noted.

In: Birk, L., Loth, G., Jotzo, L., Binder, K., & Frischemeier, D. (Eds.) (2026). *Statistics and Data Science Education in STEAM. Proceedings of the Satellite Conference of the International Association for Statistical Education (IASE), September/October 2025, Münster, Germany*. © ISI/IASE.

The Introductory Statistics unit, with a code STAT1170, is structured into five modules, each spanning two weeks. Each module was assessed by three individualised quizzes that add up to 100 marks each, from a pass-level set of questions to two more difficulty levels. All quizzes are automatically graded. The university grade categories are Fail (F, <50%), Pass (P, 50-64), Credit (CR, 65-74), Distinction (D, 75-84), High Distinction (HD, 85-100), and Fail Hurdle (FH, reduces final mark to 49).

*Hypothesis*

We expected to find that students from backgrounds outside of Australia would find the most difficulty understanding heavily accented English. Indeed, the inspiration for the project was a Hindi-speaking student who asked if it would be possible to "translate English into English" as they were having trouble understanding another professor.

Our experimental hypothesis was that full fee-paying foreign students who listened to the synthetic videos would outperform students who didn't, but that we would not see that effect among Commonwealth-funded students.

## METHODOLOGY

The lectures were pre-recorded and incrementally released on the university's learning management system week by week. In 2024 semester 2, we also released voice-changed (synthetic) lectures in parallel. Students could choose whether they wanted to listen to the original lectures, the synthetic lectures, or both. They were also informed about the study and its goals and given the option to opt out completely. No student did so, although many opted out by default by ignoring the synthetic lectures.

*Transcription and voice cloning process*

The lecturer from the original lectures was available, and we used Eleven Labs' fast voice cloning. He read a paper out loud for a few minutes, we recorded it and supplied the audio file to Eleven Labs. The result of the text-to-speech model was assessed by the native English speaker on the project, who assessed it as having "surprisingly little of the lecturer's original accent" and "a more native-like speed and cadence."

We took each lecture recording and processed it through a Whisper-based speech-to-text engine (V3 large). This created a transcript with timestamped segments. Normally, Whisper can segment at sentence boundaries, but the original lecturer's unusual pauses and timing meant that many of the segments were in the middle of phrases. These partial segments were merged. There were also artifacts and mistakes in the transcript --- Whisper models often hallucinate words (introduce extraneous words) in silent sections. These were reviewed by hand, and a native English speaker corrected the lecturer's grammar. Towards the end of the study, when the native English had more of a handle on the kinds of errors the lecturer made, ChatGPT was used to find errors and correct them.

This transformation is a sample in Table 1. Corrections made before audio rendering.

Table 1. Corrections made before audio rendering.

| **Version** | **Duration** | **Text** |
|---|---|---|
| Transcript of original Lecture | 10.00s | *<u>And</u> previously we used confidence interval to estimate the mean IQ score among the rural and remote school children.* |
| Corrected text which was rendered to audio using a clone of the lecturer's voice | 8.54s | *Previously, we used <u>a</u> confidence interval to estimate the mean IQ score among the rural and remote school children.* |

The corrected texts were passed segment-by-segment into the Eleven Labs text-to-speech model, creating an audio file for each segment. We wrote a program that concatenated these audio segments together at the positions of the original audio segments, removed the audio from the original lectures, and replaced it with the synthetic audio stream.

*Data analysis*

At the end of the semester, we extracted the students' enrolment type from the university's admission system. With only a few exceptions, all students were either Commonwealth-funded (domestic) students- who are much more likely to be native English speakers since it implies Australian citizenship- or full-fee-paying (international) students who are much less likely to have English as their native language. There are a small number of other possible enrolment types (such as Open University) that were included as part of aggregated numbers without being individually analysed.

The Macquarie University Human Research Ethics Committee has approved the study (HREA-17409). The dataset was de-identified prior to analysis to ensure compliance with ethical and privacy standards.

*Ridge regression*

We extracted and performed a regression analysis against the number of videos of each type they watched, grouped by enrolment type, attempting to predict their final grades in the unit. While simple linear regressions are frequently used in education studies, we opted for more contemporary statistical techniques robust to distributional assumptions. Ridge regression controls overfitting by shrinking predictor coefficients and stabilising estimates and handles multicollinearity (correlation between predictors). The regularisation parameter is chosen via cross-validation to balance bias and variance. Like standard regression, the coefficients indicate the relative importance of predictors. We maintained the standard frequentist p-value threshold of 0.05 and made no Bonferroni correction since the experiments we perform are in line with our pre-registered data analysis.

For all 3 regressions, we performed a Ridge (linear) regression and used leave-one-out cross-validation to find the optimal alpha parameter. That is, we found A, B and C that minimised

$$\text{LOOCV Error} = \sum_{i=1}^{n} \left(y_i - \left(A \cdot x_{\text{original},i} + B \cdot x_{\text{synthetic},i} + C\right)\right)^2 + \alpha(A^2 + B^2)$$

Where:

- $y_i$ is the actual grade of student i,
- $x_{\text{original},i}$ is the number of original videos watched by student i,
- $x_{\text{synthetic},i}$ is the number of synthetic videos watched by student i,
- A is the coefficient for the number of original videos watched,
- B is the coefficient for the number of synthetic videos watched,
- C is the intercept,
- $\alpha$ is the Ridge penalty parameter.

We chose Ridge regression because There is some correlation between the number of original videos watched and the number of original videos watched, as shown in **Fehler! Verweisquelle konnte nicht gefunden werden.**. While it is relatively low, Ridge regression can reduce the danger of an overconfident model. We also observed that the data is dominated by students who watched very few videos, with a few influential outliers (Figure 2), and Ridge regression is less sensitive to such outliers than ordinary least squares.

Table 2. Correlation statistics.

| Enrolment type | Pearson correlation between the number of original videos and synthetic videos watched by each student |
|---|---|
| Commonwealth funded (domestic) | 0.22 |
| Full fee-paying (international) | 0.13 |

The low positive correlations suggest that students who watch original-voice videos may also engage with modified-voice videos, and vice versa. The higher correlation for domestic students indicates a

greater tendency to engage with both formats, whereas international students seem more selective in their viewing.

RESULTS

The resulting regression coefficients are shown in **Fehler! Verweisquelle konnte nicht gefunden werden.**. Our initial hypothesis that international students would benefit most from the synthetic voice videos, with no such effect among domestic students, has been disproved. It shows very little impact --- or possibly a slightly negative impact --- on the grades of international students who watched synthetic videos. What surprised us was the extremely strong effect of the grades of domestic students.

Table 3. Regression statistics.

| Enrolment Type | Grade prediction coefficient grade = A * number of original videos watched + B * number of synthetic videos watched + C | | |
|---|---|---|---|
| | A | B | C |
| Commonwealth funded (domestic) | 0.90 | **1.49** | 68.8 |
| Full fee-paying (international) | 0.90 | 0.84 | 76.12 |

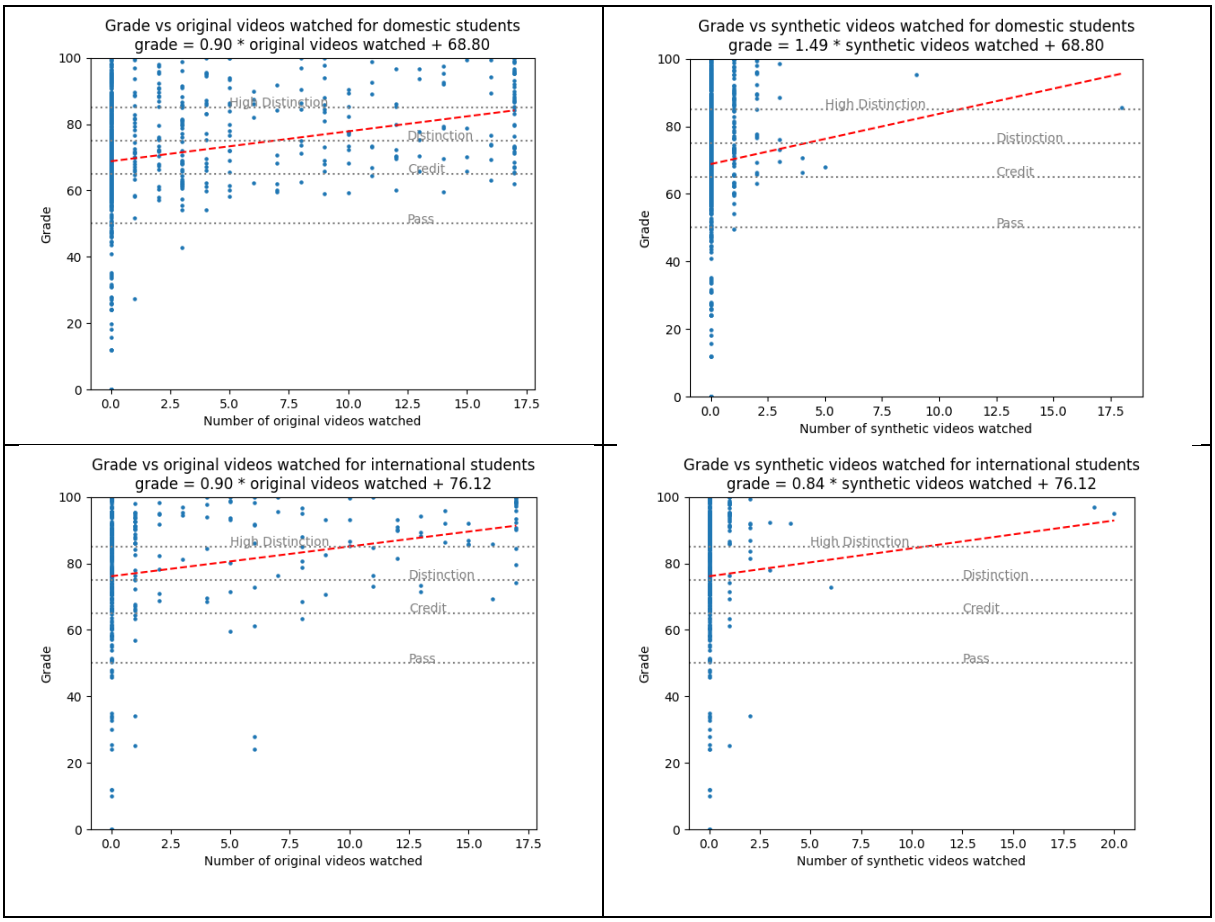


Figure 1. Plots of videos watched by a student versus grade. The first row of plots correspond to the domestic cohort, and the second to the international.

Figure 1 shows the number of videos watched, with trend lines indicating expected marks assuming zero views of the other video type; however, synthetic videos were less popular than the originals, so these results are based on a smaller subset of students.

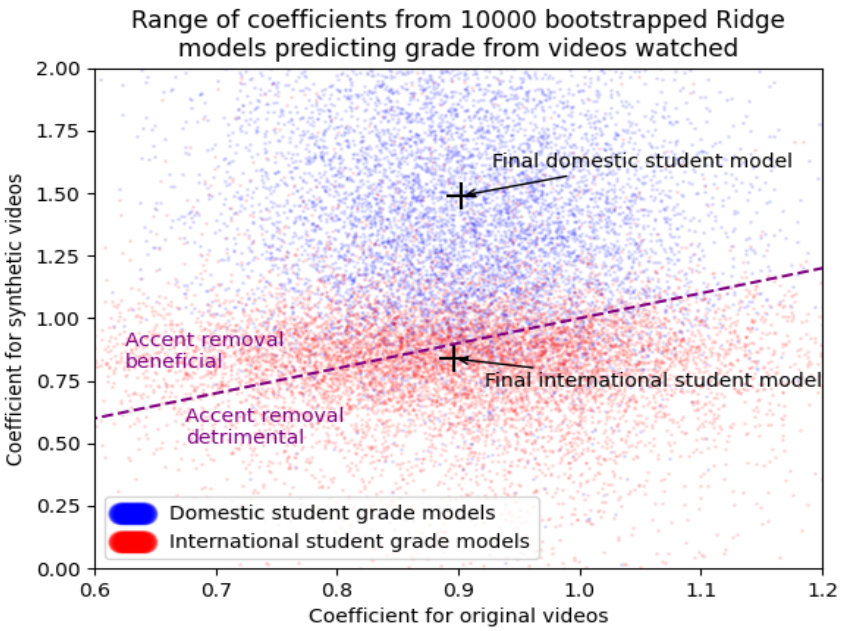


Figure 2. Visualisation of bootstrapping calculation of p-values.

*Bootstrapping p-value*

To calculate a p-value, we bootstrapped from the data 10,000 times, sampling from the domestic student data and performing a cross-validated Ridge regression on each bootstrap sample to see how often more extreme regression coefficients appeared than what we saw in the international student data, and then likewise, we sampled the international student data repeatedly training up Ridge models to see how often a domestic model had a more extreme set of values. These numbers should be broadly similar since they are symmetric.

The p-value for the international model being the same as a random selected domestic model (i.e. the proportion of bootstrapped domestic student models that were further away from the final domestic model than the international model) was 0.038. The p-value for domestic model being the same as a random selected international model (i.e. the proportion of international student models that were further away from the final domestic model) was 0.015. Both p-values are below the cut-off threshold of 0.05. A visualisation of this is in Figure 2 with visualisation of bootstrapping calculation of p-values. From this we conclude confidently that there is a difference in the responses from international and domestic students.

*Effect size*

We calculated the effect size by asking "if we were only providing students with synthetic videos, how much would their marks improve".

The difference in marks for a domestic student watching a synthetic video over an original is 1.49 – 0.90 = 0.59. The average number of videos which were watched in original format and not also in synthetic format by domestic students was 2.56, letting us predict that an average domestic student uplift from this intervention would 1.52 marks (out of 100). The standard deviation for domestic students is 17.6, meaning that there is a tiny effect size (0.09).

However, this is mostly because students don't watch the lecture videos at all. Out of the 764 domestic students in the course, 485 watched zero synthetic videos and zero original videos. There are interventions that we could do that would help these students but fluency improvements in videos that they never planned to watch is obviously not going to be one of them.

If we ask the question about the effect size of the students who do watch videos, it is somewhat more impressive. Among the students who watch lecture videos, the average number of original videos watched (where they didn't also watch the synthetic video) is 7.40. Thus, we could expect those students to get 4.32 marks higher, giving an effect size of 0.25 – small, but not insignificant.

## DISCUSSION

Extrapolating the per-video uplift to the maximum possible benefit, a domestic student who listened to all 22 synthetic videos could be expected to lift their grades from a pass to a credit, a credit to a distinction, a distinction to a high distinction. This seems to be an extraordinary result beyond all credibility, although informal conversations with students who are native English speakers suggested that many find it completely believable that having a non-native lecturer could cost them a whole grade level.

Nothing that we did in this study could not also have been done by an enterprising and lateral-thinking student wishing to lift their own grades. The tools we used are within the budget of many students acting alone; a group of students acting together would be able to afford this relatively easily. It does not require any institutional support – if the student has access to a video of a lecture, they could transcribe and regenerate it themselves. Thus, even if every university decided to establish a moratorium on the use of fluency improvement technology, that would still not prevent the technology from being used.

Eleven Labs charged us $360 USD for the four months we worked on this project, as we never ran through the included quota in any month. For this analysis, we ignored the cost of creating the software to replace the audio, as it will hopefully be amortised over its use in many units over many future semesters.

Spread among the 764 students, this intervention cost only slightly more than $0.47 per student, with the cost per effect size (calculated as dollars per student per standard deviation) a mere $5.46 (it is common to give the price per student per 0.1 effect size, i.e. $0.54). This would make it one of the most cost-effective tertiary education interventions. Even with a class size ten times smaller, it would still be an outstandingly cheap intervention. If the effect size and experiment stand up to replication --- as we hope they will --- it will be very difficult for any university to justify not deploying this intervention.

### *Considerations and limitations*

We acknowledge that this is a delicate and sensitive topic that touches on issues of identity, equity, power dynamics, and inclusion. We do not want to stigmatise lecturers operating in languages other than their native language, but at the same time, we have a duty to deliver the best results for students. Homogenisation of accents is also a kind of cultural takeover; but if a lecturer is keen on delivering quality outcomes for their students, who are we to say that they cannot use modern technology to achieve that outcome? On the other hand we also don't want to put pressure on non-native lecturers to adopt similar modifications, since that would raise ethical concerns about authenticity and self-expression. Nor do we want to discourage diversity (linguistic or otherwise) in academic hiring.

The high number of students (485 out of 764) who watched zero videos raises questions about selection bias – perhaps the students who watched the synthetic videos were unusually curious students who would have done extremely well anyway. Perhaps it is not the fluency improvements that lead to the improved results, but that highly curious students do well in introductory statistics and are also more likely to explore listening to synthetic-voiced videos.

The p-values reported are not as strong as we would like to see. Replication of this study is urgent and important, and can be done without significant funding or major effort. All that is required is a unit of study with lectures delivered by a non-native speaker, where the lectures are recorded, and the audience includes students who are native speakers.

### *Future research and open questions*

Our intervention has two parts: the correction of the lecturer's grammar, and the regeneration of their voice in a more neutral accent. Is it possible that only one of these is required for the intervention to be effective?

Why is the effect only seen in domestic students? One of our colleagues (who proudly describes himself as a "fluent speaker of broken English") has observed that native speakers have more difficulty understanding broken English than non-native speakers do, as they have less experience of hearing it than a non-native speaker does: most non-native speakers have heard a lot of broken English from their classmates in English class. He suggested that the effect we saw is related to the additional cognitive

load the domestic students (who will have less experience of hearing broken English) are under. Is he correct in this hypothesis?

What about languages other than English? For example, do fluent German speakers (for example) perform better when listening to lectures with synthesised German voices than with a non-native German speaker?

If students only have access to fluency-improved videos, do they change their viewing habits? Future research should explore:

1. The distinct impacts of grammar correction versus accent modification.
2. The long-term effects on student learning and engagement.
3. The broader implications for international faculty and cultural diversity in academia.
4. Alternative approaches to enhancing lecture comprehensibility.

*Conclusions*

This study takes an initial step towards understanding how voice technology can support educational outcomes. For educators and researchers in statistics and data science, especially within STEAM settings, these findings show that making lecture delivery more accessible through voice technology could remove a subtle but significant barrier to understanding, particularly when teaching technical material that relies on building conceptual knowledge. From a pedagogical perspective, the findings indicate that AI technologies can help increase engagement and reduce obstacles to learning technical content delivered through complex or unfamiliar mathematical language. While our results show promising possibilities, they also highlight the need for a balanced approach that values both effective teaching and cultural diversity in university environments. Future implementations should carefully consider how to improve student understanding while maintaining the authentic voices and valuable perspectives that international staff bring to higher education.

REFERENCES

Banerjee, A. V., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukerji, S., Shotland, M., & Walton, M. (2016). Mainstreaming an effective intervention: Evidence from randomized evaluations of 'Teaching at the Right Level' in India. *MIT Department of Economics Working Paper No. 16-08. SSRN.* http://doi.org/10.2139/ssrn.2846971

Ben-Zvi, D. (2016). Three paradigms in developing students' statistical reasoning. In S. Estrella, M. Goizueta, C. Guerrero, A. Mena, J. Mena, E. Montoya, A. Morales, M. Parraguez, E. Ramos, P. Vásquez & D. Zakaryan (Eds.), *XX Actas de las Jornadas Nacionales de Educación Matemática* (pp. 13–22). SOCHIEM. https://www.sochiem.cl/documentos/actas-jnem/2016-valparaiso-xx-pucv.pdf

Binkowski, K. P. (2023). Mastery of learning - does it make a difference to students' online engagement and performance in a first-year statistics unit?. In E. M. Jones (Ed.), *Fostering Learning of Statistics and Data Science - Proceedings of the Satellite conference of the International Association for Statistical Education (IASE).* ISI/IASE. https://doi.org/10.52041/iase2023.107

Croke, K., & Atun, R. (2019). The long run impact of early childhood deworming on numeracy and literacy: Evidence from Uganda. *PLOS Neglected Tropical Diseases, 13*(1). https://doi.org/10.1371/journal.pntd.0007085

McClure, K. L., & Chen, H.-T. M. (2024). "I could not understand anything they said!": Non-native English-speaking instructors, online learning, and student anxiety. *Psychology of Language and Communication, 28*(1), 233–260. https://doi.org/10.58734/plc-2024-0010

Sabharwal, M. (2011). Job satisfaction patterns of scientists and engineers by status of birth. *Research Policy, 40*(6), 853–863. https://doi.org/10.1016/j.respol.2011.04.002

Thurn F. (2023, January 25). Enhancing 'teacher presence' to improve online engagement. *TECHE Macquarie University's learning and teaching blog.* https://teche.mq.edu.au/2023/01/enhancing-teacher-presence-to-improve-online-engagement/.

Zhao, G., Ding, S., & Gutierrez-Osuna, R. (2021). Converting foreign accent speech without a reference. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29*, 2367–2381. https://doi.org/10.1109/TASLP.2021.3060813