English Programmes

—

English without Borders Evaluation Summary

30 June 2024

English without Borders-EwB- is a strategy developed by the British Council (BC) that seeks to strengthen the learning of English of students through a virtual English course with an intensity of between 4 and 6 hours per week. The classes are led with a predominantly communicative, participatory, and student-centred approach. The British Council has implemented the English without Borders program since 2021. During its first year, it was offered to students from public schools in Bogotá. In 2022, the program was extended and was offered to students in public schools in the cities of Barranquilla, Bogotá, and Bucaramanga. In 2023, it focused again on public school students in Bogotá and opened two new lines for adults. The first one was for public school English, primary, and high school teachers who were interested in either strengthening or learning English. The second one, which is of interest to the present executive summary, named Todos a la U, was for adults who live in the city of Bogotá and are interested in strengthening or learning English while also attending technical training in one of four Universities in the city. This last project was designed and implemented for the Agencia Distrital par ala Educación Superior, la Ciencia y la Tecnología-ATENEA.

Methodology

Assigning students to groups per Cohort in Todos a la U-ATENEA

The number of available seats per Cohort was determined based on the agreement between the British Council and ATENEA regarding the total number of groups and levels. Students were assigned to groups based on their level, which was defined by their core score on the EnglishScore test, and their preferred schedule. Students who did not present the EnglishScore test were automatically assigned to the Entry level, and those who failed to submit a preferred schedule were assigned to groups with available seats.

A sample of students from Cohort 1 was chosen to present the KITE (Kaplan) test; they were an additional evaluation sample for the project. However, due to the low response rate, the licenses were given to any student who chose to participate and respond to the exam.

Earlier in the process the British Council and Universidad de los Andes agreed to complete the observation process only in Cohorts 1 and 4 (due to the size and the timing of the classes per Cohort). A sample of groups were chosen to be observed considering variables such as: having two observations per teacher, from groups in different levels, and in two hour-long classes. However, the process later required different adjustments.

The classes to be observed in Cohort 1 were recorded by the Evaluation Center; this meant that there was always an external individual who would join the class and record it from beginning to end. The recordings were then uploaded to a secure platform created and managed by the Evaluation Center, where the coders (external experts trained by the Evaluation Center for the purpose of the observation) would be able to log in and watch the classes. The groups and sessions that would be observed were previously chosen (using the variables mentioned before) by the Evaluation Center. The classes from Cohort 4 were recorded by the teachers themselves using Zoom. This meant that there were no external visitors during the session but also that the characteristics of the recordings depended on the Zoom settings available to the teachers. It also

meant that teachers who did not want to be recorded would not record the class. The selection of the groups and sessions to be observed was made considering the total number of sessions recorded and their characteristics. The total number of sessions observed per Cohort is as follows: in Cohort 1, a total of 188 sessions were observed (recorded over a two and half weeks period), and a total of 200 sessions were observed in Cohort 4 (recorded in one week).

In both Cohorts, given the large number of groups in each one, an agreement was made to adjust the observation process. In order to make more observations, instead of watching the entire class, the coders would limit themselves to the first 50 minutes of the class (excluding breakout groups) and the last 5 to 10 minutes.

All classes were coded only by a single observer, but these were professionals who had completed the observation training process previously. Results were then processed to create scores per scale (the instrument is divided into scales reporting on planning, interactions, methodology, use of language, and monitoring and feedback) and presented per city and overall to highlight the percentage of observed classes that attained a high, medium or low score per scale.

The first two Cohorts of the Todos a la U-ATENEA program began implementation in the first semester of 2023. The last two began during the second semester of 2023 and ended in the first semester of 2024.

Candidates for the program had to express interest by submitting a form and presenting some documentation. ATENEA would review their applications and inform the universities which candidates would become students in the program. Initially, all were required to participate in the English classes offered as part of the training. Therefore, in order to determine who was assigned to which English level, all students were required to present the EnglishScore test and to select their preferred schedule option(s).

The assignment was then made considering the total number of groups per Cohort required by ATENEA and the British Council (per their agreement), the students' scores and schedule preferences, and the remote teachers' availability and scores. Exceptions had to be made; for instance, students who did not present the English Score test would be automatically assigned to the Entry level, but if they had some knowledge, then the remote teacher would assess their level and recommend the program to move them to a class in the correct level. Also, if students did not inform their preferred schedule, they were assigned to their corresponding level in a group that still had seats.

Because the definition of how many groups will be formed assumes that all groups will have 25 students, challenges appeared during the assignment because most of the students are initially assigned to the Entry, Fundamentals, and Progress levels, mainly the first two. Also, the demand for specific schedules is higher than others and may not always match the availability of the candidates for remote teachers. Because of this last point, new rules were required. For instance, having a minimum of 10 students to open a group, establishing a maximum of 30-35 students per class, or moving students to a lower or higher level when their core scores are close to the cut score. Another challenge was completing the assignment in steps because the program was still missing candidates or information about candidates. This happened primarily in Cohorts 1 and 2.

The following are the original results of the assignment process:

- Cohort 1: 272 groups, 7,137 candidates were assigned.
- Cohort 2: 128 groups, 2,840 were assigned Cohort 3: 201 groups; 5,455 candidates were assigned
- Cohort 4: 349 groups; 7,568 candidates were assigned

Class observation in Todos a la U-ATENEA

Since the program began in 2021, a class observation exercise has been conducted in all lines of the strategy to assess whether the quality of the implementation adhered to the standards set by the British Council. This helps inform the British Council on how to improve the delivery of the classes and how to train and support remote teachers.

Since the implementation in 2021, the sessions that will be observed have been recorded to ensure coders have access to the information with the necessary detail to report what happens in each one. By recording the sessions, the information that is being shared on the screen, the gestures of the remote teacher, the participation via the chat, the participation of the students, and the additional documentation shared can be easily reviewed by the coder.

Over the years, changes have also been made to improve the instrument, considering how the strategy has evolved. The changes have been made in agreement with the British Council to adjust questions or response options and make the report more objective. An example of this is the introduction of tallies since 2022 or the changes in response options that will make them more accessible for the coder to understand and use.

In the case of ATENEA, an agreement with the British Council was made to make observations in some groups of Cohorts 1 and 4. Because of the number of groups per Cohort, an additional adjustment was made: not to observe the entire class session but a significant portion, which would be informative enough to enable the coder to report how the remote teacher leads it. Also, an agreement was made to exclude breakout rooms from the observation process because they have important differences in how the remote teachers use them and how they can be observed (or not) depending on the permissions granted by the remote teacher and the Zoom recording settings.

The report submitted to the British Council explains how the sessions to be observed were chosen, recorded, and observed. As previously agreed with the British Council, descriptive data offers insights into how the materials are used and the overall results obtained with the questions that serve a descriptive purpose. For the questions that are scored, the results are presented per aspect (the instrument has five aspects that are scored: lesson plan, class interactions, methodology, use of English, and monitoring and feedback), showing the percentage of sessions that attained a score that allocates them in a low, middle or high level. The high level means that the implementations fulfil the expectations set by the British Council, while the low indicates that there is considerable room for improvement.

Implementation monitoring

The monitoring of the English without Borders program in 2023 and 2024, allowed to track specific indicators and results related to the implementation of the program in 2023 and 2024 as part of the Todos a la U program. This approach allowed for measuring changes in students' English proficiency and provided insights into which characteristics, either of the participants or the program itself, were correlated with these results. However, it's important to note that observed results may not necessarily be directly attributable to the program as the implementation monitoring does not measure the impact of the program on the target population.

Monitoring Design:

- Monitor implementation (inputs, activities, and outputs) and results over time.

- Allows to compare improvements in outputs and outcomes with targets.

- Allows to identify which (participant or program) characteristics are associated with better outcomes.

- However, observed results cannot necessarily be attributed to the program because there is no point of comparison or counterfactual.

Data collection methodology:

Data was collected through several digital instruments and surveys:

- English proficiency was obtained via pre and post-test (proficiency at baseline and after the course ended) using the English Score instrument.

- Survey of student characteristics, attitudes, and satisfaction with the course
    - Data was collected by British Council staff
    - Administrative data on student attendance was also considered in the analysis.

- Data analysis methodology:

- Comparison of student test scores at the endline with their baseline scores.

- We used the survey to identify patterns in score changes by various student characteristics.

- Analysed barriers to attendance and performance by attendance rates.

- It is important to note that the selection process for students in the control and treatment groups was not randomized, meaning that differences between the two groups were expected. To account for this, a Propensity Score Matching methodology was applied to students from Cohorts 4.1 and 4.2 to make the groups more comparable.

- However, not all students assigned to the English course actually enrolled, resulting in partial participation. Another way to interpret the results is by comparing those who participated in the course to those who did not, regardless of their initial assignment. Considering this, and to adjust for self-selection, an instrumental variables strategy was employed, focusing only on students who adhered to their original assignment.

Scope of the monitoring:

- Results are based on 9,052 students who took both entrance and exit exams (39.35% of the students that were part of the program)

Impact evaluation

Evaluating a program is particularly challenging when all students are eligible to participate (there is no control group) because there is no point of comparison. As this was the case for English without Borders in the Todos a la U program in 2023, one potential solution was employing an encouragement study, which involved boosting the attendance of certain students via text messages. Through randomly assigned messages, an encouragement strategy was implemented to improve attendance among selected students who participated in Todos a la U in 2023, aiming to estimate the impact of English without Borders on English learning for those who increased their attendance due to the messages.

The feasibility of this evaluation is based on generating a notable difference in participation levels to achieve sufficient statistical power for estimating the effect of attendance on English proficiency. By comparing participation rates between those who received messages and those who didn't, causality can be attributed to the program. However, in our encouragement study, messages did not have a sustained impact on attendance rates. This can be attributed to the British Council's tracking system, which regularly contacts students, potentially diminishing the effectiveness of additional WhatsApp messages, and some other students' characteristics beyond the scope of the program. Given these findings, conducting the program's impact evaluation through an experimental motivation design was not possible.

Based on these findings, we decided to stop sending messages and pursue a different evaluation approach. In 2024, an impact evaluation of the English without Borders program was conducted on the English proficiency of the Cohort 4 participants of the Todos a la U program. Cohort 4 had 46,759 applicants, divided into two sub-cohorts: Cohort 4.1, which included the 31,200 students who applied before the end of August 27, and Cohort 4.2, which comprised the remaining students who applied after that date.

Students in Cohort 4 could apply for spots in technical programs such as gastronomy, Business Process Outsourcing (BPO), and other fields. Additionally, they could apply to study English through the English without Borders program offered by the British Council. If they applied to more than one technical program, they had to specify a priority order for their choices. All students from Cohort 4.1 who selected gastronomy or BPO as their first priority and also applied for the English course were granted a spot in the English without Borders program. These 3,641 students form our treatment group. Out of these, 2,332 students (64%) chose to participate in the English course. However, due to limited spots, students from Cohort 4.2 who made the same selection (choosing BPO or gastronomy as their first priority and also selecting English) were admitted to the technical programs but were not offered a spot in the English course. These 2,298 students form our control group.

It is important to note that the selection of students between the control and treatment groups was not random, so we expect that the two groups are different. Hence, a Propensity Score Matching methodology was applied between students from Cohort 4.1 and 4.2 to make the two groups comparable. After matching by score and individual characteristics, the results showed

that those who were offered a spot in the English without Borders program program had higher English learning scores than those who did not receive the offer. The difference was close to 19 points, equivalent to an increase of about 9% over the control group average.

However, not all individuals assigned to the English course actually took it, resulting in incomplete participation. Another way to analyse the results is by comparing those who actually took the English course with those who didn't, regardless of whether they were assigned to take it. Out of the 757 people assigned to the control group who took the test, 22 participated in the English course (roughly 3%). Among the 1,030 people in the treatment group who took the test, 183 did not take the course (about 17.8%).

To correct for self-selection, an instrumental variables strategy was used, which considers only those who adhered to their assignment. That is, students from Cohort 4.1 who were assigned to take the course and did so, and those from Cohort 4.2 who were not assigned to take the course and did not take it. The instrumental variables methodology indicates that 79.3% of the sample behaved according to their assignment (this is commonly referred to as the complier population). The local average treatment effect results (LATE) show that for compliers, there was a 27.7-point increase in test scores (13.4% relative to the average of those who did not take the course).

Evaluation questions:

- Class observation: Were the classes implemented with the quality and standards the British Council expected?

- Implementation monitoring. The key question here was "How much did students' English knowledge improve after participating in the English without Borders course as part of Todos a la U? How does this learning vary according to individual characteristics or attendance?"

- Impact evaluation. The key question here was "Did the English without Borders course of Todos a la U impact the English language abilities of students? If so, what was the size of the effect?"

Scope of the evaluation:

The instrumental variables strategy assumes that: 1) there are no individual characteristics correlated with the instrument that affect the treatment, and 2) the instrument does not directly influence the outcomes except through the program. This would imply that being assigned to the treatment (the instrument), compared to Cohort 4.2, which was not assigned, affects English proficiency levels only through the likelihood of taking English without Borders, and that being part of Cohort 4.1 instead of 4.2 is not correlated with other individual characteristics. If this holds true, we can use this instrument to estimate the effect of taking the program by being assigned to the treatment and use an instrumental variables strategy. With this estimator, we are calculating the local average treatment effect (LATE) on individuals who comply with the instrument.

An alternative approach is to apply the propensity score matching strategy and match those who actually participated in the treatment with those from the control group. While this strategy is similar to the previous one, the assumption here is stronger, as it assumes that individuals who participated in the program and those who did not are identical once matched by

observable characteristics. If this assumption holds, we would be estimating the Average Treatment Effect on the Treated (ATT). However, since there is a self-selection issue, it becomes more challenging to argue that the control group individuals are identical, as we don't know who among them would have taken the program if it had been offered. Therefore, we must be more cautious in interpreting this estimation as the causal effect.

Limitations

Class observation:

- The way the classes are being recorded strongly impacts the observation process. When the class is recorded by a third party that visits the classroom and uses another software (different from Zoom), features like the chat, the list of participants, and the gallery format are used, and this enables the observer to understand better how many students participate in the class and how elements like a compensation strategy are being used or how compelling is the teacher's invitation to participate or speak in English. When teachers record their own classes, and the speaker mode is activated, then only the person who is talking is seen, and the observer has a hard time understanding how big the audience is and how students react to the tasks and strategies.

- Another limitation to consider in the process is that the sessions observed in Cohort 1 were closer to the end of the intervention. From the remote teachers, we learned that new variables affected how the program was delivered. For instance, most participants preferred the late afternoon schedule because they were not in classes or at work; however, this hurt participation because most students would be in transport during the class and could not share screens, log in to the Richmond platform or open their microphones to participate. By the end of the implementation, most students could not attend classes because of the beginning of a new semester at university.

- Because there are fewer groups in the higher levels of the program, the observation strategy has more observation in the levels of Entry, Fundamentals, and Progress. It would be interesting to further explore the dynamics of groups in higher levels.

Implementation Monitoring:

- Students were required to complete four instruments or surveys (pre and post-test on English proficiency, a survey on student characteristics and attitudes, and a satisfaction survey) over the six-month program period. This likely impacted the response rates of the instruments and surveys, with only 12% of students completing all four.

- Results are not causal estimates and are not likely to be representative of all students (students who attended more were probably more motivated).

- Students who took both exams attended 76.7% on average, while those who took baseline test only 20%.

Impact Evaluation in 2023:

- Messages were not successful in increasing attendance for two reasons:

- The British Council regularly communicates with students through calls and texts to boost attendance. The effect of additional messages was marginal when students received several other messages during the course.

- Student absenteeism seems predominantly linked to structural factors like unpredictable work schedules or family obligations, rather than a lack of interest. This implies that reminders may have limited effectiveness, especially for the most vulnerable students.

- Without a significant and meaningful difference, conducting the program's impact evaluation through an experimental motivation design was not feasible.

- Based on these findings, we stopped sending messages and pursued an alternative evaluation approach to be implemented in 2024.

- Other methodologies are being considered to study the effect of the program since the result is not that the program is ineffective but that the method used did not allowed to have a conclusion.

Impact Evaluation in 2024:

- The instrumental variables (IV) strategy relies on two assumptions: 1) no individual characteristics are correlated with the instrument that affect treatment, and 2) the instrument only influences outcomes through the program. In this case, being assigned to the treatment (Cohort 4.1) affects English proficiency levels only by increasing the likelihood of participating in the English without Borders program. If these assumptions hold, the IV approach can estimate the local average treatment effect (LATE) on individuals who comply with the assignment.

- Alternatively, the propensity score matching method can be used to compare those who actually participated in the treatment with those in the control group. While this approach is similar to the IV strategy, it involves a stronger assumption: that participants and non-participants are identical once matched by observable characteristics. However, due to self-selection, it's harder to argue that control group members would have participated if given the chance, making this estimate less straightforward to interpret as a causal effect.

Evaluation findings

Key findings

The main findings of the **class observation** in 2023 and 2024 were:

- Overall, in Cohort 1:
  - Lesson plan: 18.6% of the sessions observed achieved a high-level score, while 45.2% attained a low lever score.
  - Use of English:42.6% of the sessions observed achieved a high-level score, while 35.6% attained a low lever score.

- Overall, in Cohort 4:

- Lesson plan: 27.5% of the sessions observed achieved a high-level score, while 37.5% attained a low lever score.
- Use of English: 29.5% of the sessions observed achieved a high-level score, while 32.5% attained a low lever score.

The main findings of the **implementation monitoring** in 2023 and 2024 were:

- The results of 9,052 participants (39.35% of the initial participants) were analysed, showing that those who took both the pre and post-tests had an average attendance rate of 76.7% and an average increase in scores of 34.4 points after six months of class.

- In the post-test, 66% of students who were at below A1 level in the pre-test progressed to A1, while 33% of those at A1 level in the pre-test advanced to A2. These transitions represented the highest advancements across CEFR levels.

- Women's learning difference between pre- and post-test was on average positive, statically significant (pre-post English score test difference: 34.3, $p < 0.01$) and no different than that of males. There are also significant and positive results in underrepresented groups such as Afro-descendants or indigenous individuals (pre-post English score test difference: 39.3, $p < 0.01$), people with disabilities (pre-post English score test difference: 34.9, $p < 0.01$), and LGBTIQ community (pre-post English score test difference: 33.4, $p < 0.01$). However, these results are not statistically different from those of non-Indigenous individuals, people without disabilities, or those who are not part of the LGBTIQ+ community.

- Women had higher attendance rates than men (79.6% vs. 72.2%), and individuals from higher socioeconomic strata also attended more frequently (81.4% vs. 76.2%). Additionally, those with tertiary education showed greater attendance compared to those with only a high school diploma (81.95% vs. 78%).

The main findings of the **impact evaluation in 2023** were:

- Up to week 8, messages failed to significantly increase student attendance in class. Despite a slight increase of one-third of an hour (0.33) in attendance for the treatment groups compared to the control group, this difference is not statistically significant.

- Messages did not have a sustained impact on attendance rates due to the British Council's frequent communication efforts to boost students' attendance, and the prevalence of structural barriers to student attendance (i.e., time scarcity, job scheduling conflict), suggesting limited effectiveness of reminders, particularly for vulnerable students.

The main findings of the **impact evaluation in 2024** were:

- Participating in the English without Borders course improved English proficiency levels in students' part of Cohort 4 of Todos a la U program.

- An effect ranging from 18.6 to 27.6 points on average was found.

- The average score on the EnglishScore test was 218.6, with a standard deviation of 110.2. This indicates that students who took the course improved by between 8.5% and

12.6% on the test. This change represents between 16.8% and 25% of the standard deviation.

- These effects are notable not only due to their statistical significance but also their size (compared to other educational interventions).

- Results are relevant considering that the English without Borders course is not very intensive, is delivered virtually, and the participants face challenging conditions that make attending courses difficult.

Overall, in Cohort 1:

- Lesson plan: 18.6% of the sessions observed achieved a high-level score, while 45.2% attained a low lever score.

- Class interactions: 19.7% of the sessions observed achieved a high-level score, while 52.7% attained a low lever score.

- Use of English:42.6% of the sessions observed achieved a high-level score, while 35.6% attained a low lever score.

- Monitoring and feedback: 9.6% of the sessions observed achieved a high-level score, while 56.4% attained a low lever score.

Overall, in Cohort 4:

- Lesson plan: 27.5% of the sessions observed achieved a high-level score, while 37.5% attained a low lever score.

- Class interactions: 31.5% of the sessions observed achieved a high-level score, while 68.5% attained a low lever score.

- Use of English: 29.5% of the sessions observed achieved a high-level score, while 32.5% attained a low lever score.

- Monitoring and feedback:6.5% of the sessions observed achieved a high-level score, while 58.5% attained a low lever score.

Moreover, based on the satisfaction survey and students who took both the pre and post English proficiency test, the Net Promoter Score (NPS) is 70.68% with the following distribution:

- Promoters: 62.17%
- Detractors: 8.51%
- Passives: 29.33%

Lessons learned

Class observation:

- Adapting the observation scheme to observe sixty minutes of the class focusing on the first fifty minutes (excluding break out groups) and the last ten minutes, proved useful.

Coders reported that the first fifty minutes allowed them to understand the dynamics of the class and have enough information to answer the questions.

- It is crucial to provide coders with precise instructions on the minutes they need to observe, including when to skip a breakout room or a coffee break. This clarity is essential to prevent them from losing track of time and to facilitate their observation of specific minutes.

- While it may be more convenient for remote teachers to record their classes using Zoom, the platform's features and settings can pose challenges for coders. It is important to consider these challenges and explore alternative recording methods that allow for a more thorough observation of class dynamics.

- There is a higher number of groups and students in the Entry, Fundamentals, and Progress levels. There are fewer groups and students in Connections, Evolution, and Specialist. Therefore, most of the observations are made in the first three levels. An interesting element in the process is that in Cohort 4, a few teachers had several groups in the Progress and Connections levels. Therefore, because each teacher would be observed twice, the number of observations possible at those levels was reduced by having a few teachers offering the classes.

Implementation monitoring and impact evaluation:

Main Achievements (Good Practices):

- Students with high attendance rates (>70%) identified the following factors that helped them stay engaged in the course:

- Effective Teaching Methods: Students praised the dynamic and engaging teaching methods employed by the British Council team, which kept them motivated and interested.

- Supportive Learning Environment: The supportive and encouraging atmosphere created by the teachers fostered active participation and facilitated learning.

- Relevance and Enjoyment: The program's content and activities were perceived as relevant and enjoyable, contributing to sustained interest and attendance.

What Worked Particularly Well and Why:

- In the satisfaction survey, students with high attendance rates (>70%) highlighted the following program success factors:

- Engaging Teaching Methods: The use of diverse pedagogical tools and activities made classes enjoyable and effective, enhancing learning outcomes.

- Supportive Teachers: The dedication and professionalism of the teachers, coupled with their ability to create a positive learning environment, significantly contributed to student engagement and satisfaction.

- Relevant Content: The program's content was perceived as relevant and valuable by students, addressing their needs and interests effectively.

What Didn't Work So Well and Why:

- Limited Response Rates: Both surveys' response rates were relatively low compared to the total cohort size, indicating potential challenges in capturing information from all students. This could be due to the time of the year when the surveys were administered (November-December, close to the Christmas holidays). Integrating the data collection in the class schedule and making sure that instructors assume responsibility in data response rates may help. This limited response rate may affect the program's ability to gather comprehensive insights and address areas for improvement effectively.

- Work-related Absences: While work commitments were cited as the primary reason for class absences, this factor is beyond the program's control. Strategies to accommodate students with work obligations may need further exploration to enhance attendance rates.

- Perceived Lack of Usefulness: While only 0.3% of students reported finding the class not useful, addressing the reasons behind this perception could further improve program effectiveness. Understanding and addressing factors contributing to perceive a lack of usefulness can enhance student engagement and satisfaction.