Measuring Student Mindsets at Scale in Resource-constrained Settings: A Toolkit with an Application to Brazil during the Pandemic

Guilherme Lichand[1], Elliot Ash[2], Benjamin Arold[2], Jairo Gudino[2], Carlos Alberto Doria[3], Ana Trindade[4], Eric Bettinger[4] and David Yeager[5]

Abstract: Accumulating evidence that a growth mindset – the belief that intelligence is not fixed and can be developed – is associated with better educational outcomes has spurred research on the topic, from how to measure student beliefs more accurately to how to promote challenge-seeking behavior in the classroom. Most of this research is, however, concentrated in high-income countries, as measuring and intervening on student mindsets typically relies on connected devices, which are often unavailable in low- and middle-income countries' schools. This paper develops a toolkit to measure student mindsets in resource-constrained settings. Concretely, we adapt the computer-based survey instruments to text messages in the context of Brazilian secondary public schools. Collecting mindset survey data from 3,570 students in São Paulo State as schools gradually reopened in early 2021, we validate our methodology by matching key patterns in our data to previous findings in the literature. We also train a machine learning model on our data and show that it can (1) accurately classify students' SMS responses, (2) accurately classify student mindsets even based on text written in other media, and (3) rate the fidelity of different interventions to the growth mindset concept.

[1] Department of Economics, University of Zurich; guilherme.lichand@econ.uzh.ch.

[2] Department of Political Science, ETH Zurich.

[3] Department of Economics, University of Brasília.

[4] Graduate School of Education, Stanford University.

[5] Department of Psychology, University of Texas at Austin.

1. Introduction

A growth mindset is the belief that intelligence is not fixed and can be developed (Dweck, 2006). Such psychological trait might influence how adolescents react to challenges at school, as well as their disposition to learn new concepts. Student mindsets have been shown to be empirically associated with math and language grades, especially among the poor; enough to temper the association between income and academic achievement (Claro, Paunesky and Dweck, 2016). Interventions that foster a growth mindset have also been shown to causally improve adolescents' educational outcomes (e.g., Rege et al., 2020; Yeager et al., 2019), with effects concentrated in schools where fixed mindsets are most prevalent (Yeager et al., 2022). Recent work has now begun to document how student mindsets interact with moderators such as classroom culture, e.g., as expressed through teachers' discourses (Hecht et al., 2023a, 2023b).

Despite the promise of studying mindsets to gather insight into what factors lead some students, unlike others, to display beliefs that intelligence is malleable and into which schools and students could benefit the most from such interventions, most evidence on student mindsets comes from high-income countries (in particular, the United States). This is problematic for two reasons. First, survey instruments designed to capture student mindsets might not work as intended in low- and middle-income countries At least part of the reason is that mindset data is typically collected through computer-based surveys (and growth mindset interventions, typically delivered as part of the same session), while schools in resource-constrained settings rarely have access to computers and the internet (Brossard et al., 2021). In Brazil, the setting of our study, only 41% of schools had at least one functioning computer, and only 19% had internet speed above 50Mbps (TIC Educacão, 2020).

This is even more challenging if researchers and practitioners are interested in tracking student mindsets in settings under conflict – where any face-to-face data collection is challenging –, or in forced migrants or refugee populations, who might not be part of the formal school system even where computers would have been available. The second reason why the scarcity of evidence from low- and middle-income countries is problematic is that interventions designed to promote a growth mindset might not work as intended in such settings. This is because they might not be attuned to the local context or, even if they are, because contextual factors impose additional constraints that preclude mindset change or, else, do not lead to better educational outcomes even conditional on a growth mindset. As a leading example, Ganimian (2020) finds that a growth mindset intervention designed after its US counterpart failed at changing student mindsets in Argentina.

This paper develops a toolkit to measure student mindsets in low-resource settings. Concretely, we surveyed 92,234 public school students (grades 6-12) over text-message (SMS) in São Paulo State, Brazil, between February and April 2021, as schools were gradually reopening to in-person activities in the aftermath of the Covid-19 pandemic. In this context, not only are computers seldom available in these schools, but also, face-to-face surveying would have been impossible due to contact restrictions during that period.

We also train a machine learning model on our data and show that it can (1) accurately classify students' SMS responses, (2) accurately classify student mindsets even based on text written in other media, and (3) rate the fidelity of different interventions to the growth mindset concept.

Our results show that it is possible to track student mindsets even in low-resource settings as long as context-appropriate survey instruments and technologies are used. The paper also paves the way for context-appropriate growth mindset interventions, by making all datasets, dictionaries, and codes available for future use by other researchers

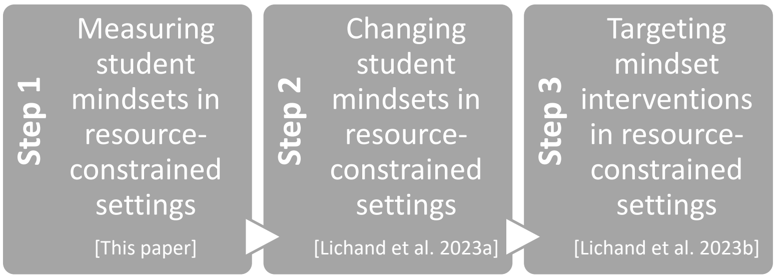
2. Framework for this research program

This study can be understood as the first of three steps in a research program on *measuring and supporting student mindsets in resource-constraint settings* (see Figure 1). In the first step, we develop instruments to accurately measure student mindsets in the absence of computer-based surveys. As articulated in Section 5, this toolkit should be relevant even when phone-based surveys are not feasible, thanks to the classifier algorithm that we make available as part of the replication package for this paper.

Once we can measure student mindsets in resource-constrained settings, the second step is to intervene to convert fixed mindsets into growth mindsets – i.e., breeding beliefs that intelligence is malleable, such that a student can always improve relative to him or herself. To ensure interventions are context-appropriate and cost-effective, it is essential to be able to measure their impacts. We take on this task in a first companion paper (Lichand et al., 2023a), which randomly assigns 12 different SMS-based growth mindset interventions across Brazilian secondary public-school students – the same population as the one in this study. The paper evaluates the impacts of these interventions on student mindsets (based on the analyses undertaken for this study) and educational outcomes. We illustrate the potential of this connected research program by using our trained classifier to rate the 12 interventions when it comes to their fidelity to the growth mindset concept (see Section 3.4).

Last, once we have evaluated the extent to which different interventions can promote a growth mindset among students in a resource-constrained setting, the third step in this research program is to target each student with the message most likely to both change their mindset and improve their educational outcomes as a result. This is what we take on in a second companion paper (Lichand et al., 2023b), which uses machine learning models to predict the optimal intervention for each student, and then documents the extent to which optimal targeting can make a difference for improving educational outcomes, and whether it can be achieved in practical applications – even with limited data, typical of resource-constrained settings.

Figure 1 – Framework for 'Measuring and Supporting Student Mindsets in Resource-Constraint Settings' research program



Step 1

Measuring student mindsets in resource-constrained settings

[This paper]

Step 2

Changing student mindsets in resource-constrained settings

[Lichand et al. 2023a]

Step 3

Targeting mindset interventions in resource-constrained settings

[Lichand et al. 2023b]

Notes: Elaborated by the authors.

3. Methods

This section starts by describing how we adapted the computer-based interventions and survey instruments to the SMS format in Section 3.1, followed by how we surveyed students over SMS, in Section 3.2. Section 3.3 then describes how we classified student responses and how we validated our methods using secondary data and focus group discussions. Last, Section 3.4 describes the machine learning methods used to automatically classify SMS responses into mindsets, and beyond.

3.1 Adapting the growth mindset intervention and survey instruments

3.1.1 Computer-based intervention, survey instruments, and the need for adaptation

The methodology for developing psychological interventions proposed by Yeager et al. (2016) set the stage for many growth mindset studies after it. Given the effectiveness of this intervention in promoting a growth mindset for 9th grade students both in the US (Yeager et al., 2016; Yeager et al., 2019) and Norway (Bettinger et al., 2018), our study was designed to extend and test this intervention in Brazil, a developing country where secondary education is compulsory but less than 80% of students graduate from primary education and less than 60% graduate from secondary education.

The computer-based intervention has two versions (Yeager et al., 2019): 1) a treatment session, which discusses the malleability of intelligence, explaining the importance of productive struggle and of making mistakes, based on the science of the brain – in particular, highlighting the metaphor that the brain is ‘like a muscle’ (“the harder one works, the stronger it gets”) –; and 2) a placebo session, which merely illustrates how the brain works, focusing on its different anatomical parts and the extent to which different functions are or are not associated with specific brain regions. Students are typically randomly assigned to one out of two versions. As part of the session, in both versions, students have to write letters motivating other students to participate in future sessions. We include illustrations of the introductory screen, of the treatment version, and of the prompts asking students to write these letters in Appendix A.

Such letters can be used to classify student mindsets, along with closed-ended questions that elicit the degree to which participants agree with four statements related to the malleability of intelligence:

(1) “You have a certain amount of intelligence, and you really can’t do much to change it.”
(2) “Your intelligence is something about you that you can’t change very much.”
(3) “Being a ‘math person’ or not is something that you really can’t change. Some people are good at math and other people aren’t.”
(4) “When you have to try really hard in a subject in school, it means you can’t be good at that subject.”

Participants rate each of these statements in a Likert scale (in our application, the scale has 6 points, from strongly/somewhat/weakly disagree to weakly/somewhat/strongly agree). In most applications, subjects are considered to have a growth mindset if they disagree (at least weakly) with all the four statements. Such mindset measures have been used and validated in numerous studies, demonstrating that they strongly predict grades and performance on behavioral tasks.

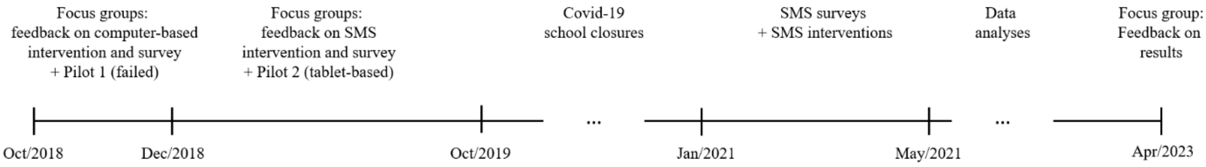
We anticipated two main challenges. First, that we would be unable to implement a computer-based intervention, given the severe infrastructure challenges faced by Brazilian public schools. Second, that at least some of these mindset questions would have to be rephrased to ensure they would be understood by our target population, and that they would effectively capture the psychological construct we were interested in. As the following subsections demonstrate, both concerns proved to be correct, and ultimately led us to not only adapt the intervention and survey instrument to a different format and technology, but to write up this paper to document the outcomes of these adaptations, as well as what we learned throughout the process.

3.1.2 Local partnerships and timeline

The process of adapting the intervention relied on partnerships with Rio de Janeiro and São Paulo States' Secretariats of Education. Ranking among the largest student populations in Brazil, these Secretariats navigate complex educational challenges. The São Paulo State Secretariat alone manages a public school system with more than 40,000 students per cohort, scattered across 1,000 schools.

Figure 2 outlines the timeline of our field activities. We started the project in Rio, where we set out to pilot the computer-based intervention and corresponding survey instruments over the last school quarter of 2018. Sections 3.1.3 and 3.1.4 describe the activities prior to this first pilot, and the reasons for why it ultimately failed – which ended up being instrumental for the transition to SMS that ensued. In 2019, we were able to pilot both the computer-based intervention (with the help of tablets) and the SMS interventions and survey instruments in Rio; Sections 3.1.5 and 3.1.6 describe the implementation details and the pilot results.

Figure 2 – Timeline


Focus groups: feedback on computer-based intervention and survey + Pilot 1 (failed)
Focus groups: feedback on SMS intervention and survey + Pilot 2 (tablet-based)
Covid-19 school closures
SMS surveys + SMS interventions
Data analyses
Focus group: Feedback on results
Oct/2018
Dec/2018
Oct/2019
...
Jan/2021
May/2021
...
Apr/2023

Notes: Elaborated by the authors.

In 2020, in face of political transitions in Rio (which made it difficult to keep our study moving forward in the State), we moved the study to São Paulo. We developed the partnership during the period of school closures and rolled out the SMS-based study over the first two school quarters of 2021. Section 3.2 describes how we surveyed students throughout that period. Section 3.3 then describes how we classified student responses, analyzed data, and validated our results using different methods – including focus group discussions conducted in 2023, described in Section 3.3.3.

3.1.3 Focus groups prior to piloting the computer-based intervention

We started the adaptation process with feedback from teachers and parents instead of students, following the translation protocol developed in Bettinger et al. (2018) for Norway. Such feedback was elicited through several focus groups over the last school quarter of 2018 and the first two school quarters of 2019. The goal of these focus groups was to leverage the adults in the community and their knowledge of the target population of students to identify potential issues with the translation of abstract concepts from English to Portuguese and collect suggestions for changes.

In the interest of initiating the adaptation process and collecting information as quickly as possible, we opted for a convenience sampling procedure. Based on recommendations from the Rio de Janeiro State Secretariat of Education, we contacted the most responsive school representatives to intermediate the recruiting of school principals and teachers for this stage of the adaptation process. Two schools agreed to support the initial focus groups and interviews, and eight schools agreed to participate in the pilot.

We were able to recruit eight readily available teachers at one of the partnering schools for the initial focus group. These teachers expressed excitement with the intervention and its potential, voiced concerns about the language and level of abstraction required by the material presented, and suggested modifications to make the language more accessible to the targeted student population. Unfortunately, due to the low engagement of parents with the schools in this context, we were only able to interview two mothers that came to one of the schools to provide feedback on the content developed for the SMS delivery format. While one of the mothers had no difficulty understanding and engaging with the short messages, the other one had significant difficulty and had to have the meaning of each text explained in more than one way.

With students, we were able to conduct a total of six focus groups, with 7-8 students each. One group examined the placebo content from the computer-based material and helped rewrite the growth mindset instrument in a way students could better relate. Three groups examined the content for computer-based intervention, and the remaining two were responsible for examining the intervention content in text message format.

Overall, enumerators leading the student focus groups reported lack of interest from students, with difficulty in obtaining substantial feedback. In particular, the length of these activities was raised as an issue, as students would lose interest quickly. However, some students did engage with the material, with mentions of "hope" that they could learn more and become better at learning, and sparked debates over how to work on it.

Based on this feedback, however thin, we refined the Portuguese version of the computer-based intervention script. In particular, we included additional visual elements (see Appendix A) to make it more appealing to our target age group (12–17-year-olds). An additional adaption was to feature voice-overs of all text in each screen, for a variety of reasons: including students with special needs, connectivity issues that could get on the way of loading images, and engaging students with low reading proficiency (which are not uncommon in Brazil, even as late as in secondary school). Our hope was to further refine the intervention script following the pilot results.

3.1.4 Piloting the computer-based intervention

The first pilot study was carried out at the end of the 2018 school year, with a pool of 880 students and parents and 23 teachers across 8 schools. However, only 438 students, nine teachers, and 201 parents completed the baseline survey – and even then, with a significant delay. Together with the limitations in equipment and unreliable information about mobile phones, this made it impossible to collect outcome measures for this pilot.

The implementation challenges faced while running this first pilot revealed two of the major challenges related to the Brazilian public education setting. First, the requirement of opt-in parental consent in this pilot proved to be a significant barrier to recruitment, with an enrollment rate of only 32% of students. Anecdotal evidence offered by school staff suggested that parental engagement with the schools was very low in general, as many of the parents felt unable to contribute or participate in their children's education because they help multiple jobs and had very little to no formal schooling. Second, schools in this context lacked the basic equipment needed for the computer-based intervention. When available, the only computers with internet access in these schools belonged to the administrative staff office, so only a couple of students could do the computer-based activity at a time.

3.1.5 Adapting the intervention and survey instruments to SMS

In agreement with the Secretariat of Education, we made three major revisions to the implementation protocol to enable a second pilot study. First, we replaced parental opt-in consent with an assent and opt-out protocol for the parents. Second, we would bring in our own tablets, such that we could effectively implement the computer-based intervention at schools. Third, we would adapt the intervention and survey instruments to text messages (SMS), such that we could have an alternative media that did not rely on school infrastructure to be rolled out for a subsequent large-scale randomized control trial.

Text messages were appealing because this capability comes pre-installed in every phone (no need for smartphones), it does not require a data plan (or even airtime, with reverse billing for replies at no cost for participants), it cannot be uninstalled, and its push notifications cannot be turned off. Data plans are expensive in Brazil, where a minimum airtime top-up is of the order of USD 6. As a result, a recent report indicates that Latin America displays one of the lowest connectivity levels in the world (PWC, 2022). While almost 100% of people have online access at some point during the month, such access is not perennial – averaging only 12 days of connectivity throughout the month (PWC, 2022). That

As we discussed the need to transition to SMS, the Secretariat of Education committed to requiring parents to update their mobile phone information upon enrollment every year, since parents' phone numbers was either missing or outdated to a great extent, restricting the sample of students that could be assigned to the SMS intervention.

Adapting the growth mindset intervention to the SMS format was a lengthy and complex process, led by Brazilian EdTech Movva[6]. The main challenge was incorporating the key messages from the hour-long computer-based intervention script, filled with rich imagery and supported with voice-over audio, into the SMS format – which is very limited in nature, without capabilities for audio or image, and constrained to 160 characters per message. Thankfully, Movva had extensive experience in running communication with Brazilian public school students and their families over SMS, and set out to adapt the content of their communication to promote the essential elements of the growth mindset messages present in the computer-based script (see Lichand et al., 2023a, for additional details).

The pilot SMS intervention featured 2 messages per week over the course of 7 weeks, targeting the phone numbers informed as part of the informed consent to participate in the pilot study.

[6] Guilherme is a co-founder and partner at Movva.

The computer-based intervention featured a baseline session, whereby we could measure student mindset before they were exposed to either treatment or placebo content, and a an end-line session, conducted approximately a month later.

When it comes to the survey questions, initially we used the same four mindset questions translated to Portuguese for the computer-based intervention, sending them all at the end of the 7 weeks of the pilot intervention. Unlike the computer-based session, however, participants did not have to rate their degree of agreement with each statement in a 6-point scale; rather, they merely had to state whether they agreed with each statement ('yes' or 'no').

Results would highlight the need for adaptation of the questions as well, as the next subsection makes it clear.

3.1.6 Pilot results

This pilot study was conducted during the first semester of the 2019 school year. We were able to obtain consent from eleven new schools to participate in this pilot, for a total of 636 participating students. Participants were randomly assigned to either the placebo or treatment session of the computer-based intervention, and later randomly assigned to either the SMS intervention or no SMS communication (the control group). Interventions were cross-randomized, such that we could analyze their results separately.

Results revealed a positive and significant effect of both the computer-based and SMS interventions on two out of the four growth mindset questions. Surprisingly, one of the mindset questions ("When you have to try really hard in a subject in school, it means you can't be good at that subject") was associated with a negative impact of both the computer-based and SMS intervention, and we did not find a significant effect associated with either treatment on the remaining growth mindset question.

By aggregating these items into an index, we found suggestive evidence of an overall positive effect for both interventions, with a 0.115 s.d. increase in the probability of displaying a growth mindset for the SMS intervention, and a 0.163 s.d. increase for the computer-based one. Nevertheless, focusing on students that started off with a fixed mindset (measured in the baseline tablet session), we found a divergence between the two formats, as the computer-based was associated with a positive impact on the growth mindset index, while the SMS intervention showcased a negative impact.

Although mixed, these results were encouraging and helped shed light on the potential for improvement of the growth mindset intervention. Thus, we decided to reiterate the adaptation process with new focus groups and a small-scale pilot for the SMS before implementing this intervention at scale.

3.2 Surveying students over SMS in the main study

After extensively piloting the survey questions with our target population to ensure that students could understand the questions and that their answers were consistent with the theoretical constructs, preventing potential mismatches between the instruments and the sociocultural context of the global south, we arrived at the final survey instrument adapted to the SMS format of our main study, as follows:

(1) “Do you agree that if you need to study a subject very hard this means that you are not good at it?”
(2) “Do you agree that intelligence is a fixed trait?”
(3) “Do you agree that doing well or not in math is something that you cannot change?”
(4) “Do you feel unmotivated to study after you receive negative feedback?”

Unlike the computer-based intervention, participants did not have to rate their degree of agreement with statements on a Likert scale. Instead, the statements were rephrased into questions, and participating could respond to the prompts in any way they found meaningful. If that potentially creates challenges for manually classifying responses into mindsets, it also potentially creates opportunities to leverage large language models to automatically classify student mindsets using a richer set of inputs.

Participating students received one question a week over SMS, following the order above. All participants received the same questions each week, and questions rotated every 4 weeks, over a total length of 8 weeks.

The study sample comprised 92,234 students (grades 6-12, covering mainly 12–17-year-olds) eligible to participate in the growth mindset intervention evaluated in Lichand et al. (2023a). All participants received an activation message at the very beginning of our study, explaining that they could opt out of participating in the study by texting a stop word (‘cancel’, ‘stop’, ‘exit’) at any point throughout the study.

While high-school students received SMS questions directly on their phone on record, that was not always the case for middle-school students: we targeted them directly whenever their own phone was on record, but often had to target their primary caregiver’s phone instead. Whenever the caregiver was asked instead of the student, we changed the framing of the question to make it relational; concretely:

(1) “Do you agree that if your child needs to study a subject very hard this means that s/he is not good at it?”
(2) “Do you agree that your child’s intelligence is a fixed trait?”
(3) “Do you agree that your child doing well or not in math is something that s/he cannot change?”
(4) “Does your child feel unmotivated to study after s/he receives negative feedback?”

Even though, in that case, mindsets might be reported only indirectly by caregivers, we still attribute answers to student mindsets.

All text messages were sent by Movva, in collaboration with the São Paulo State Education Secretariat, in order to securely access students’ and caregivers’ phone numbers in compliance with the General Data Protection Regulation.

3.3 Classifying and validating results

3.3.1 Hand-coding SMS responses into mindsets

We start by coding SMS responses manually. This was manageable because of the fairly low response rates (~1%), leading to 5,700 responses (4,534 non-empty and directly related to the survey question; see Section 4.1). Hand-coding was done independently by two research assistants. Inconsistencies were resolved by a principal investigator.

We classified each valid response as either a fixed mindset, if its text expresses agreement with the question statement, or a growth mindset otherwise. For students with only one valid response over the course of the 8 weeks of data collection, we classified their mindset based on their sole answer. For those with multiple valid responses, we assigned them a fixed (growth) mindset if they (dis)agree with all statements, or an undefined mindset if they provide conflicting answers (following Claro et al., 2016).

3.3.2 Validating growth mindset measurement over SMS using previous findings

We first validate our results by contrasting key patterns of the data collected over SMS relative to those documented with face-to-face surveys with students in Chile, reported in Claro et al. (2016).

Concretely, we compare (1) the wealth gradient of growth and fixed mindset; i.e., how the distribution of student mindsets changes as household income increases in each study; and (2) the wealth gradient of math and language grades conditional on student mindsets; i.e., how the distribution of student grades changes as household income increases, separately for those classified as growth or fixed mindset, in each study. In particular, we are interested in whether a growth mindset (as captured through our SMS surveys) also tempers with the association between family income and grades, just as in Claro et al. (2016).

A challenge of these validation exercises is that we do not have data on household income for the participants of our study. Instead, we use an imputation procedure. We take advantage of a different study containing both data on income brackets and student characteristics for 9th-graders in São Paulo State public schools, collected in 2016 by Lichand et al. (2022). Since both our study and that previous study collected data on student gender, race, math and Portuguese attendance, and math and Portuguese grades, we can predict the income bracket for each student using a Poisson model joint with multiple-imputation methods based on these characteristics, trained in the 2016 dataset. Importantly, report card grades not included in the imputation procedure when analyzing whether mindset tempers its relationship with grades, to avoid circularity.

3.3.3 Validating growth mindset measurement over SMS through focus group discussions

The second way in which we validate our results was through a focus group discussion in April 2023. We recruited three subjects who did not participate in the study, but whose profile matched its target population (enrolled in São Paulo State public schools in 2021, at the time of the study).

The focus group elicited their feedback on the survey instruments we used in the study, on using text messages as the media for data collection, and on the main empirical patterns we documented. For the latter, we first asked them to guess the association between gender and mindset, that between age and mindset, and that between mindset and math grades in the data. Then, we showed them the actual results, and asked if they thought they made sense based on their previous discussion.

3.4 Using machine learning to train a classifier model

We use natural language processing (NLP) methods to classify student SMS into fixed or growth mindset. Concretely, in this study we rely on a BERT (Bidirectional Encoder Representations from Transformers) classifier.

BERT is a natural language processing model that is based on a pre-trained neural network architecture. We choose BERT because it is pre-trained on large amounts of text data and can learn more nuanced relationships and representations of language as compared to various other NLP classifiers, such as logistic regression, k-nearest neighbors algorithms, and SVM (Support Vector Machine). The ability of BERT to capture the contextual meaning and relationships between words in a given sentence leads to high accuracy and performance on a variety of NLP tasks including SMS classification.

Our analyses build on student SMS responses, student letters, the script of the computer-based intervention and the scripts of the SMS interventions evaluated in Lichand et al (2023a).

We discuss three applications of that model in the subsections below.

3.4.1 Classifying SMS responses

First, we use our BERT model to automatically classify student mindsets based on their SMS responses. We evaluate the performance of our BERT model by comparing it to our hand-coded classification. Specifically, we assess the overall performance of our BERT model, and analyze whether the prediction performance differs by student groups, sample, and type of text.

To run our BERT model, we keep SMS with more than 3 characters in the sample. We drop all SMS that reply ‘yes’ or ‘no’ (in Portuguese). Furthermore, we drop all special characters. In some instances, we do not drop stop words, as we consider them useful in providing context in the architecture of the algorithm (details always included in the notes to the tables and figures).

As the SMS are written in Portuguese, we use a locally trained dataset in Portuguese to calculate numerical representations for each token (word).[7] We employ a BERT architecture aimed at assigning a numerical score between 0 and 1 to each SMS, capturing whether the SMS is associated to a growth mindset or not (after learning sequences of words and learning the context). We describe the different steps of our BERT architecture including fine-tuning in more detail in Appendix D.

We apply our BERT architecture to train models with different texts, including (i) only SMS messages sent by students with more than 3 (or 30) characters; (ii) only sentences of the computer-based intervention script (placebo/treatment scripts); and (iii) a combination of the two.

[7] Although the models were trained in Portuguese, the dictionaries could be easily translated for applications in other settings.

3.4.1 Rating student letters

Second, we apply our BERT model trained on the SMS responses sent over 2021 to rate student letters written by participants of the 2019 pilot study as part of the computer-based intervention.

Concretely, we split letters into sentences, and classify each sentence as growth mindset or not based on the trained BERT model. Then, each letter is rated with respect to its fidelity to the growth mindset concept, computed as the share of its sentences classified as growth mindset.

We then use predicted ratings to investigate whether the distribution of student mindsets based on these letters are consistent with the patterns captured through the tablet survey instruments – based on the 6-point Likert scale – in the 2019 pilot study.

3.4.2 Rating growth mindset interventions

Last, we apply our BERT model to rate the extent to which different interventions effectively capture the growth mindset concept. We do so by assessing 12 different SMS interventions evaluated in a companion study (Lichand et al., 2023a), based on our trained BERT model.

Specifically, we analyze every text message sent out as part of each intervention, classifying each of them into growth mindset or not. The, each intervention is rated with respect to its fidelity to the growth mindset concept, computed as the share of its messages classified as growth mindset.

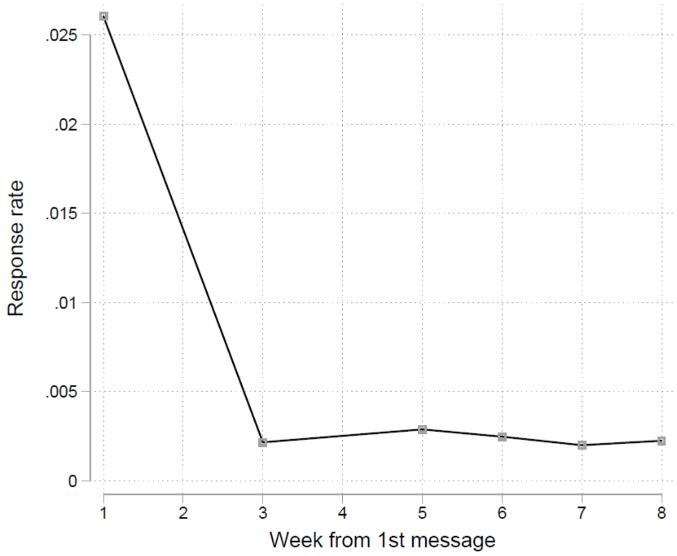
4. Results

This section starts by characterizing SMS response rates and selection into the survey in Section 4.1. Section 4.2 follows by presenting the main findings of our SMS survey (based on hand-coding SMS responses) and validating them against previous findings and focus group discussions. Next, Section 4.3 presents results for the automatic classification of SMS responses based on the BERT classifier model. Section 4.4 then compiles the results of applying that model to rating student letters, and Section 4.5, those of applying the model to rating different growth mindset interventions.

4.1 Response rates and respondent characteristics

Considering all growth mindset questions sent out throughout the study, we have 556,645 surveys over the course of the 8 weeks following the activation message. Those prompted 5,700 responses, a roughly 1% response rate. 1,116 out of those were rates as invalid (either empty or not directly related to the question prompt), leaving us with 4,534 valid answers from 3,570 unique subjects – a roughly 4% response rate (relative to the total number of participants). Out of those who provided valid responses, 85% did so only once throughout the study period. Sustaining participation over time was an even more complex challenge: as Figure 3 shows, participation quickly decays from a response rate above 2.5% in week 1 to less than 0.5% thereafter.

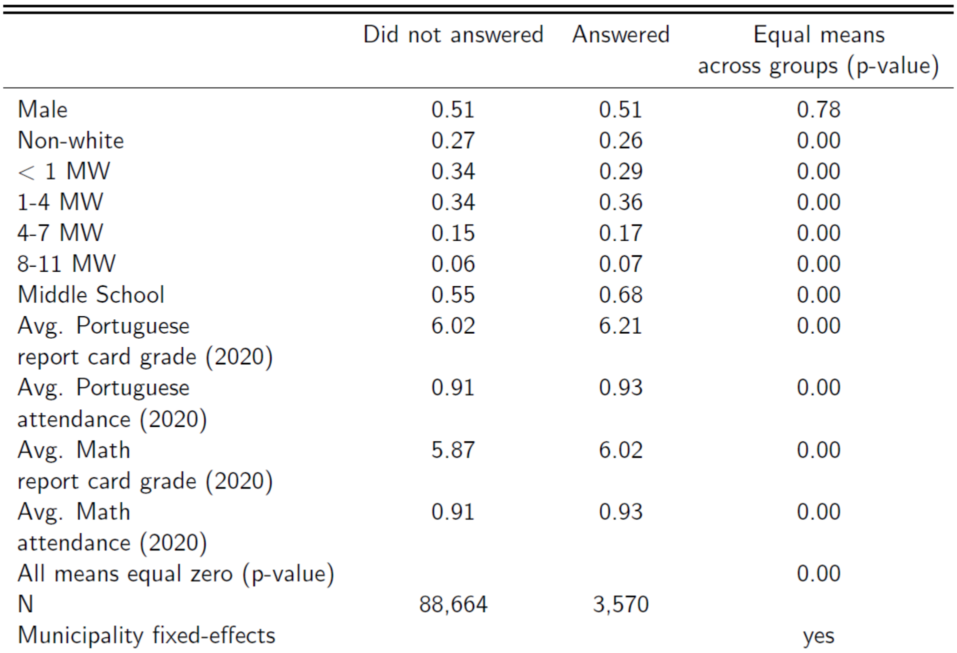
Figure 3 – Response rates over time



Notes:

A natural concern in face of low participation rates is the extent to which survey results are representative of the universe of students. Table 1 shows that, other than gender (whose distribution is nearly identical across those who answered at least one question or not), the characteristics of those who responded were systematically different than those who did not. While whites were slightly more likely to participate in the survey (74% vs. 73%), the main differences were with respect to income levels, age and school performance. 34% of those who did not respond lived in households under 1 minimum wage (based on our imputation procedure), in contrast to only 29% among those who responded. Middle-school students (or their caregivers) were also over-represented among respondents relative to non-respondents (69% vs. 55%). Last, those who respond were also significantly more likely to display higher attendance and grades in math and Portuguese.

Table 1 – Characteristics of SMS respondents relative to universe



	Did not answered	Answered	Equal means across groups (p-value)
Male	0.51	0.51	0.78
Non-white	0.27	0.26	0.00
< 1 MW	0.34	0.29	0.00
1-4 MW	0.34	0.36	0.00
4-7 MW	0.15	0.17	0.00
8-11 MW	0.06	0.07	0.00
Middle School	0.55	0.68	0.00
Avg. Portuguese report card grade (2020)	6.02	6.21	0.00
Avg. Portuguese attendance (2020)	0.91	0.93	0.00
Avg. Math report card grade (2020)	5.87	6.02	0.00
Avg. Math attendance (2020)	0.91	0.93	0.00
All means equal zero (p-value)			0.00
N	88,664	3,570	
Municipality fixed-effects			yes

Notes:

Such differences indicate that it might be important to re-weight our observations with valid surveys responses to ensure results are representative of the universe of students. It is worth mentioning that the set of students considered as the universe in Table 1 is that of those with a valid phone number on record by the São Paulo State Secretariat of Education. While that is already a selected sample (in particular, with a lower share of families living in poverty; Lichand et al., 2022), it is the relevant set for those who could have been surveyed through text messages by the Secretariat.

4.2 Student mindsets

Out of the 3,570 subjects with at least one valid answer, we classified 3,332 of them (93%) as either growth or fixed mindset; i.e., out of the 15% who answer more than one survey question throughout the study, slightly less than change the nature of their answers over time (leading us to classify them as 'undefined' mindset, following Claro et al., 2016).

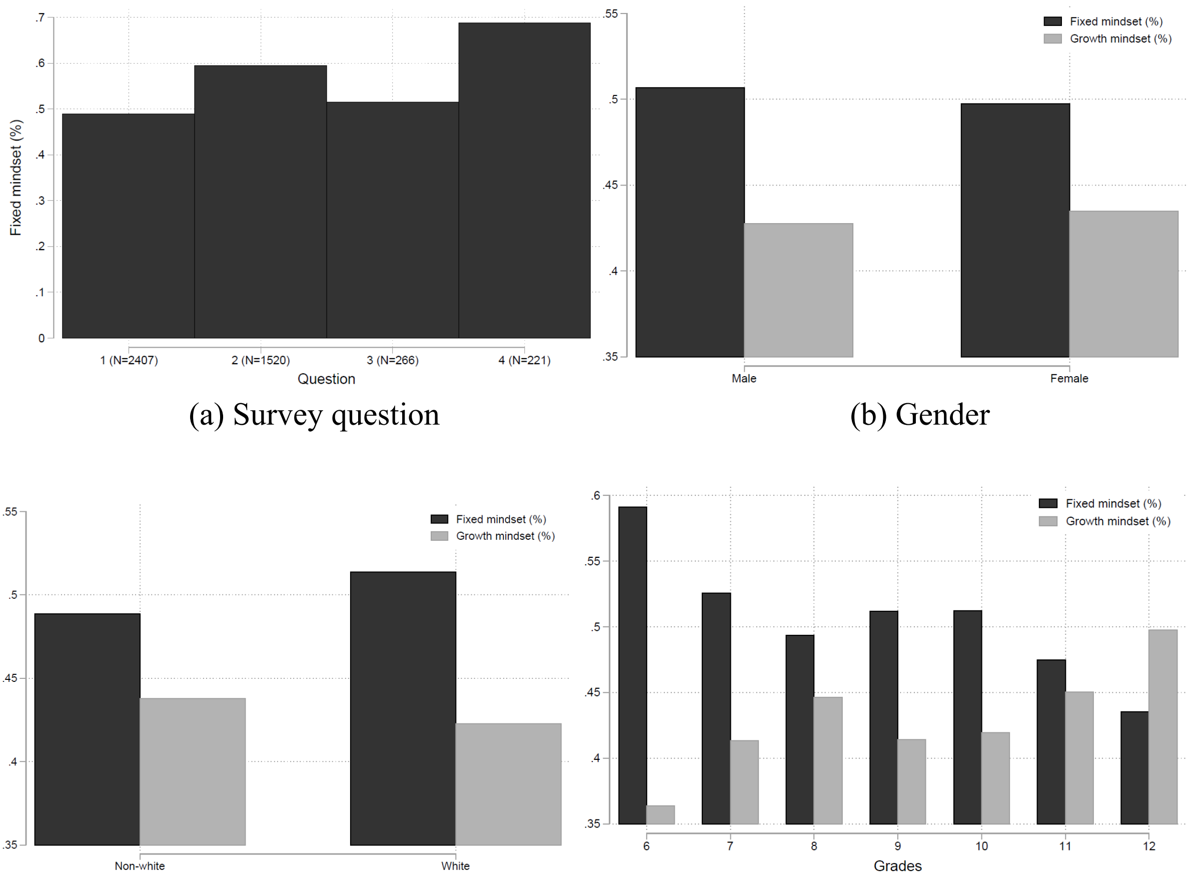
To illustrate the content of the SMS responses, prompted with the question "Do you agree that your child's intelligence is a fixed trait?", a white male 8th-grader answered "No, because we are constantly evolving, and the brain adapts to new situations. We have to use intelligence all the time to monitor these developments." We coded this answer as consistent with growth mindset. When asked "Do you agree that doing well or not in math is something that you cannot change?", a male 10th-grader of undisclosed race answered "No, if you have a good teacher you can be good at it!". We also coded this answer as consistent with growth mindset. In turn, when prompted with the question "Do you agree that your child's intelligence is a fixed trait?", a female 7th-grader of undisclosed race answered "Yes, I cannot focus on the lessons in this new normal [remote/hybrid instruction] that we are experiencing today. I prefer in-person classes", coded as consistent with a fixed mindset. Similarly, when asked "Do you agree that doing well or not in math is something that you cannot change?", a white male 8th-grader answered "Yes, math is hard. I cannot learn it", also coded as consistent with a fixed mindset.

Figure 4 showcases the distribution of SMS responses by survey question and by student characteristics (gender, race, and grade). About 50% of the responses to "Do you agree that if you need to study a subject very hard this means that you are not good at it?" and to "Do you agree that doing well or not in math is something that you cannot change?" were consistent with a fixed mindset. That share was even higher (nearly 60%) when it comes to "Do you agree that intelligence is a fixed trait?" and reached almost 70% for "Do you feel unmotivated to study after you receive negative feedback?".

Pooling all survey questions, about 50% of subjects were classified as fixed mindset, and 43% as growth mindset. That distribution is nearly identical across gender and race, even though white respondents were slightly more likely to be classified as fixed mindset (and less likely to be classified as growth mindset).

The most striking differences we documented were by grade: students at earlier grades were significantly more likely to be classified as fixed mindset. Nearly 60% of respondents at grade 6 were classified as such, in contrast to 44% of those at grade 12. In effect, the prevalence of fixed (growth) mindset nearly monotonically decreases (increases) with the school grade. While that pattern could be reconciled with a positive association between a fixed mindset and the likelihood of dropping out of school, testing that hypothesis would require collecting data from students no longer at school as well.

Figure 4 – Student mindset by survey question and by student characteristics



(a) Survey question

(b) Gender

(c) Race

(d) Grade

Notes:

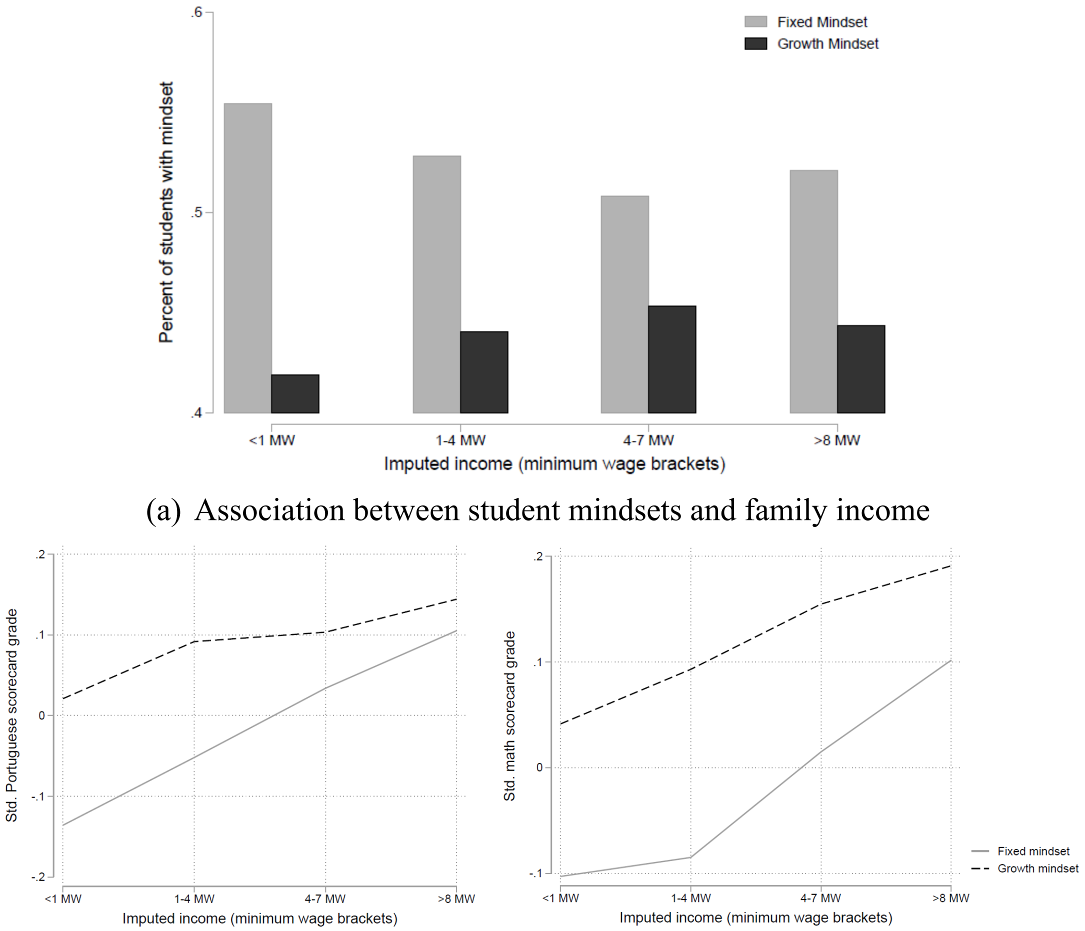
4.2.1 Comparison with Claro et al. (2016)

To validate our SMS survey methodology, we start by comparing key patterns in our data to those documented in a Claro et al. (2016), based on face-to-face surveys with Chilean students. The study used the same growth mindset questions as in the computer-based intervention, translated to Spanish – hence, directly comparable to our survey instrument.

Figure 5 displays the results. Panel (a) plots the prevalence of fixed and growth mindset by family income (based on our imputation procedure). The association between mindset and income in the SMS survey matches the patterns in Claro et a. (2016): the higher family income is, the lower (higher) the likelihood of displaying a fixed (growth) mindset. Such patterns are robust to re-weighting observation by the inverse of the predicted probability of answering the survey, based on student characteristics (see Appendix B).

Next, panel (b) plots the association between report card grades and income, separately for students with a fixed or a growth mindset (for Portuguese on the LHS, and math on the RHS). Once again in line with the evidence from Chile, results are consistent with the hypothesis that mindset tempers with the association between poverty and educational outcomes. In effect, moving from a fixed to a growth mindset is equivalent to moving from less than 1 minimum wage to at least 4-7 minimum wages in terms of standardized grades for both math and Portuguese.

Figure 5 – Validation of SMS mindset measure



(a) Association between student mindsets and family income

(b) Association between scorecard grades and income, by mindset

Notes:

Appendix B compiles additional results, namely the average association between student mindsets and report card grades, that between the former and standardized test scores, and additional robustness tests for re-weighting observations by the inverse of their probability of answering the survey (predicted based on student characteristics).

4.2.2 Considerations from focus group discussions

In April 2023, we engaged three former public-school students from the São Paulo State educational system to discuss the previous findings, along with the survey instruments and our choice of SMS for data collection media.

They generally agreed that all four survey questions are tightly associated with the concept that intelligence is malleable. Interestingly, all of them expressed agreement with at least one of the questions. This was more common for the one about negative feedback, consistent with the empirical finding that this prompt was the one associated with the highest prevalence of responses consistent with a fixed mindset.

When it comes to the choice of text messages, students were first surprised and skeptical that this was the right channel to engage students. One of them said "I never check my text messages, and I am not even sure I would know how to do it on my new phone". That could help explain the low response rates for our SMS surveys.

While they all mentioned that they use WhatsApp on a much higher frequency, they agreed that many of their peers would not have internet access every day of the month. When they learned about the fact that the Secretariat does not even have valid phone number on record for nearly 50% of the students, they all then agreed that it might be preferable to prioritize media with the highest possible reach, even if these media are analog like SMS.

When it comes to the survey results, focus group participants expected differences to be much larger in general. For instance, they expected girls and non-white students to have a much higher prevalence of growth mindset relative to boys, as a response to the social expectations from teachers and parents that make it harder for them to thrive in school otherwise. As such, they were surprised that this did not show up in the data. They also expected that having a growth mindset would be associated with a much larger difference in math and Portuguese grades than that seen in our data. Having said that, they found it valuable that our survey was able to shed light on these issues – even if to dispel misconceptions about them.

It is also worth noticing that even if a 0.2 increase in math report card grades (on a scale from 0 to 10) might be perceived as small by focus group participants, the truth is that such difference is actually quite large: as panel (b) in Figure 5 shows, that difference corresponds to more than 0.1 s.d., i.e., the typical learning rate over the course of a full school quarter.

All in all, their reactions suggest caution about over-claiming differences in the prevalence of growth mindset by income bracket. While the patterns indeed match those in Claro et al. (2016), it is true that the slope of the gradient is rather small. In turn, the association in grades is objectively large (on top of again consistent with the Chilean face-to-face survey data), providing more solid grounds for validating our survey methodology.

4.3 Classifier performance

4.3.1 Overall performance

To begin, we classify SMS sent by students into fixed or growth mindset, using a BERT model and human annotations of SMS. To measure the performance of our SMS classifier, we present a confusion matrix (Table 2). A confusion matrix is a visual representation of the performance of a machine learning model, such as our BERT classifier. It compares the ML-based predictions of whether an SMS is fixed or growth mindset with the corresponding hand-coded labeling of our human annotators. Our confusion matrix consists of four categories: true growth mindsets, true fixed mindsets, false growth mindsets, and false fixed mindsets. True growth (fixed) mindsets refer to the number of cases where the model correctly predicts a growth (fixed). Conversely, false growth (fixed) mindsets represent the cases where the model predicts a growth (fixed) mindset when the actual mindset was fixed (growth).

We use the values of the confusion matrix to calculate performance metrics of our BERT classifier. First, we calculate the accuracy rate, which is defined the percentage of correct predictions out of all predictions made by the model. Our accuracy rate is 0.81, indicating that most SMS are correctly classified. However, the accuracy rate may not give a precise representation of the model's performance, for example if the dataset is imbalanced (where the number of instances in each class is not equal). To account for this, we also show the F1-score. It is a metric that combines precision and recall, two important evaluation measures in classification tasks.[8] It is calculated as the harmonic mean of precision and recall, and provides a balanced assessment of the model's performance with regards to type 1 and type 2 misclassification errors. The F1-score ranges between 0 and 1, where 1 represents perfect precision and recall, and 0 represents very poor performance. We document an F1-score of 0.85, indicating reasonably high prediction performance.

Table 2: Confusion matrix

Actual\Predicted	**Fixed**	**Growth**
Fixed	51	22
Growth	12	97

Notes: Table presents confusion matrix of actual and predicted growth/fixed mindset SMS. Actual classification performed by human annotators, predicted classification performed by BERT algorithm. Model training was implemented based on 70/30 random split for SMS messages larger than 3 characters.

Furthermore, we present a range of additional performance measures to assess the confidence of our model. They built on the insight that our BERT classifier generates a probability distribution of SMS over their possible labels “growth mindset” and “fixed mindset”, indicating how likely BERT thinks that each label to is applicable. In this way, BERT can provide not just a single label prediction, but also a measure of confidence in that prediction, which can be interpreted as a probability. In the following, we present the overall distribution of the model's predictions, and related performance metric (ROC curve).

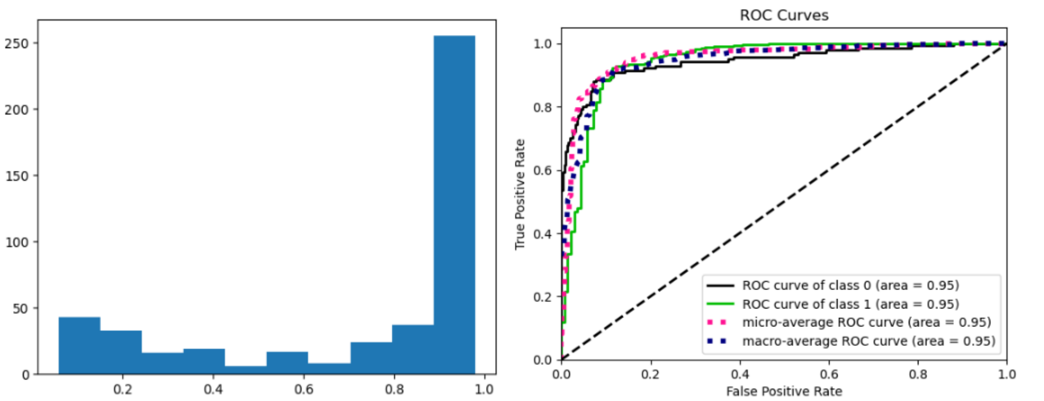
Figure 6 (LHS) showcases the distribution of predicted probabilities of being classified as growth mindset according to BERT. Specifically, the histogram bins the predicted probabilities into 10 bins, and counts the number of probabilities that fall into each bin. The large number of SMS that receive a probability greater than 0.9 indicates that our BERT model classifies SMS at high levels of confidence.

[8] Precision is the percentage of true growth mindsets out of all predicted growth mindsets, while recall is the percentage of true growth mindsets out of all actual growth mindsets.

To further assess the performance of our BERT model for different probability thresholds, we calculate a ROC (Receiver Operating Characteristic) curve. The ROC curve is created by plotting the true positive rate (TPR) on the y-axis against the false positive rate (FPR) on the x-axis for different probability thresholds. The area under the ROC curve is a commonly used evaluation metric for BERT. A perfect classifier would have an area of 1.0, meaning it achieves 100 percent TPR with 0 percent FPR, while a random classifier would have an area of 0.5, meaning it achieves the same TPR and FPR as random guessing. Hence, a high-performing BERT model will typically have a steep ROC curve that is skewed towards the upper left corner of the plot.

Figure 6 (RHS) displays the ROC curve of the BERT model, using different thresholds for the probability vector to measure goodness of fit. For different models, we find an area that is larger than 0.94. This finding, in combination with the other results of this subsection, indicates a high classification performance of our BERT model.

Figure 6: Probability distribution (LHS) and ROC curves (RHS)



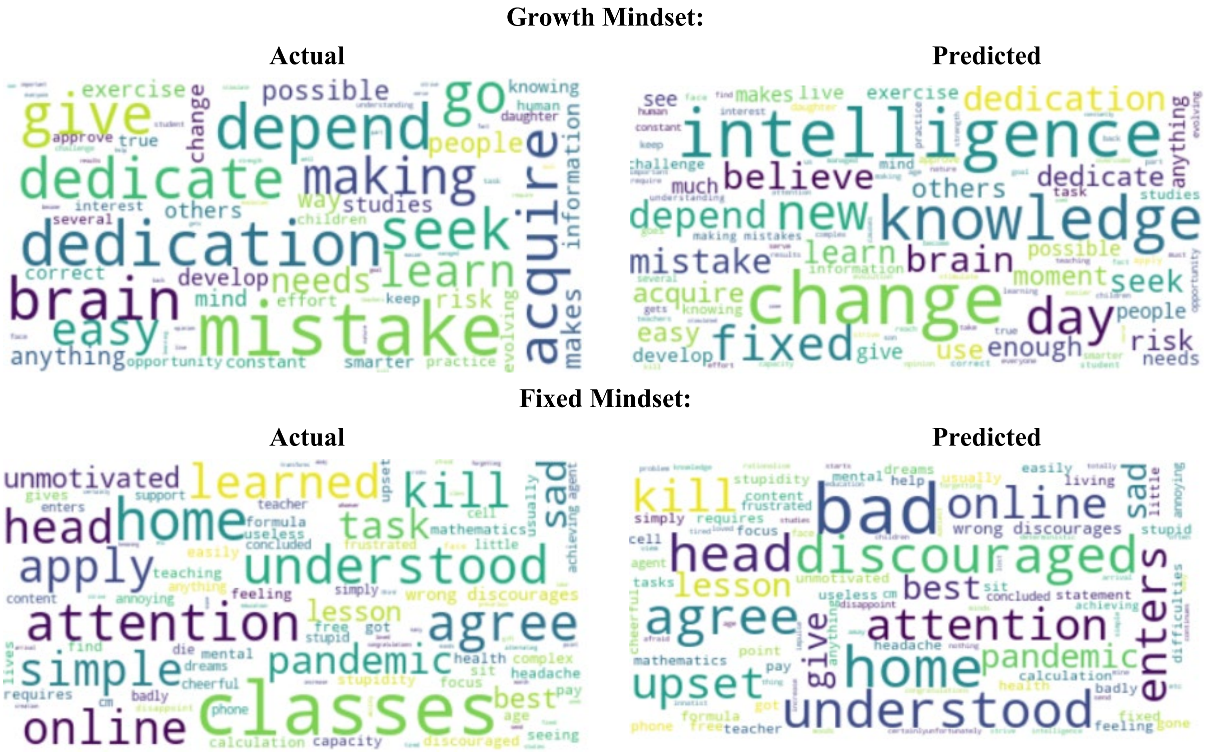
Notes: Figure on the LHS depicts histogram of the estimated probability of being a growth-mindset student, according to BERT model. Figure on the RHS depicts ROC (Receiver Operating Characteristic) curve, using different thresholds for the probability vector to measure goodness of fit. The ROC curve for class 0 represents the performance of the model in correctly classifying instances of growth mindset, while the ROC curve for class 1 represents the performance of the model in correctly classifying instances with label fixed mindset. The micro-average ROC curve aggregates the TPR and FPR across growth mindset and fixed mindset, and produces a single ROC curve. The macro-average ROC curve computes the TPR and FPR for growth mindset and fixed mindset separately and then takes the average across the two classes.

To summarize topics that students write about in the SMS they sent, and further evaluate the performance of our model, we also generate word clouds. Specifically, we present word clouds that report the most indicative words of fixed and growth mindset for the human annotator and the BERT prediction, respectively. As we are less interested in words that are common to all mindsets, we focus on relative rather than absolute frequencies.[9] As shown in Figure 7, we find that words like “dedication”, “brain”, and “change” and interestingly also “mistake” are related to growth mindset as classified by the human annotator. The corresponding fixed mindset SMS contain words like “classes”, “pandemic”, “online”, “attention” and even “kill”.

[9] In terms of absolute frequencies, the word “learn” is the most commonly used word for both fixed and growth mindsets, in the actual and predicted version (not shown).

Comparing the similarity of SMS that are most indicative of a given mindset according to the human annotator (LHS) and the BERT prediction (RHS) is another way of assessing the BERT performance. Although the words based on the predicted models somewhat differ from those based on the human annotations, they are still in line with intuitive concepts of the respective mindsets (such as "intelligence" and "change" for growth mindset, as well as "bad" and "discouraged" for fixed mindset). Furthermore, the alignment of words between the actual and predicted models within a mindset is larger if we use bigrams instead of unigrams, see Appendix Figure C1. This makes sense, as the BERT algorithm is based on learning representations through context rather than individual words.

Figure 7: Unigrams: Growth Mindset (upper graphs) and Fixed Mindset (lower graphs)



Notes: Figure shows word clouds of unigrams, by mindset (growth vs. fixed) and model ("actual" human annotator vs. "predicted" BERT) based on relative frequencies. The larger the font size, the more important is a unigram for a given mindset in relative terms.

4.3.2 Performance by subgroup

Next, we are interested in testing whether the performance of our BERT classifier differs by demographic or socioeconomic groups. Differences in the prediction accuracy across groups could potentially exacerbate student achievement gaps if SMS classification is less precise for underprivileged students, and relevant for student learning at the same time. To investigate this possibility, we merge demographic and socioeconomic characteristics to our main dataset, namely gender, race/ethnicity, and pre-intervention grades in Portuguese and math (grades in the last quarter of 2020). We then calculate the error rate, which is defined as 1 minus the accuracy rate, for the different subgroups.

As reported in Table 3, we find no substantial differences in the error rates of subgroups by gender, race/ethnicity, and pre-intervention grades in Portuguese. In contrast, the error rate is substantially lower for the best quarter of students in math (less than 4 percent) compared to the other students. In other words, math is the primary dimension of subgroup differences in classification performance, while the other dimensions are rather unaffected.

Table 3: Error rates, by selected subgroups

By Gender:

Error Rate (Female, 216 students): 0.101
Error Rate (Male, 242 students): 0.099

By Race/Ethnicity:

Error Rate (White, 242 students): 0.103
Error Rate (Mixed, 124 students): 0.096
Error Rate (Black, 11 students): 0.09

By Previous Grades - Portuguese:

Error Rate (Portuguese, Quartile 1): 0.077
Error Rate (Portuguese, Quartile 2): 0.157
Error Rate (Portuguese, Quartile 3): 0.081
Error Rate (Portuguese, Quartile 4): 0.120

By Previous Grades - Math:

Error Rate (Math, Quartile 1): 0.097
Error Rate (Math, Quartile 2): 0.090
Error Rate (Math, Quartile 3): 0.107
Error Rate (Math, Quartile 4): 0.037

Notes: Table presents error rates using BERT predictions, by subgroups of gender, race/ethnicity, and pre-intervention grades (by quartile, for Portuguese and Math, respectively). Sample size differs across subgroup analyses due to varying number of missing observations of student characteristics.

4.3.3 Performance by sample size

As mentioned in the section on methods and data, we drop SMS with fewer than 4 characters from the main sample. However, this sample still contains a number of SMS with rather limited meaningful content. Therefore, we are interested in learning how the performance of our BERT model changes if we exclude many of the rather meaningless short SMS in the training step. Specifically, we train another BERT model on a sample that keeps all SMS that contain more than 30 characters (the other sample selection criteria, such as the exclusion of special characters, remain unchanged). The training sample is now reduced by approximately 50 percent.

Overall, the performance of this reduced BERT model is largely comparable to the main model. While the accuracy drops from 0.81 to 0.79 in the small model, the F1-score increases from 0.85 to 0.87. Other measures for performance and calibration yield similar conclusions, as reported in Appendix Table C1 and Appendix Figure C2.[10]

4.3.4 Performance by type of text

The results presented so far are based on the SMS that students sent, with regards to the training the BERT model and predicting growth mindset, respectively. Now, we want to test whether training the model on another text source (that is also growth mindset related) can yield a similar accuracy for predicting growth mindset in SMS. To this end, we use the script of a computer-based intervention. There are two versions of this script, one containing text about growth mindset that a randomly selected subset of students received, and one containing placebo text that the other students received.[11]

To train this BERT model, we split each of the two texts into single sentences. First, we use our model trained on the sentences to predict the sentences themselves in a 70:30 test/validation split. This prediction works well, with an accuracy of 80%.[12] Nevertheless, the focus of this analysis lies on testing the predictive power of the script-based algorithm for the students' mindset based on the SMS. As shown in the confusion matrix in Table 4, almost all SMS are predicted to be growth mindset. The accuracy equals 59 percent. This implies that our script-based prediction approach does not work well for distinguishing the two mindsets. Apparently, mindset prediction using a given medium such as SMS requires corresponding media-specific content.

Table 4: Confusion matrix for BERT trained on computer-based intervention script

Actual\Predicted	**Fixed**	**Growth**
Fixed	2	71
Growth	3	106

Notes: Table presents the confusion matrix of actual and predicted growth/no-growth mindset for all SMS sent by students. Actual classification performed by human annotators, predicted classification performed by BERT algorithm. Model training was implemented based on the whole set of sentences of placebo (no-growth)/treatment (growth) intervention texts.

[10] To evaluate the BERT prediction performance of the small sample, we show a confusion matrix, a probability distribution, and ROC curves. To compare the BERT calibration performance between the large and small sample, we also present a calibration plot that compares the predicted probabilities of a model with the actual probabilities observed in the data. Ideally, a well-calibrated model should have predicted probabilities that match the actual probabilities, resulting in points of the calibration plot falling close to the diagonal line. Figure B3 presents a calibration plot for the large sample (blue) and the small sample (orange), respectively. The points of the large calibration model are a little closer to the diagonal line, with one notable outlier at the mean predicted probability of 0.67.

[11] The placebo text is not specifically related to fixed mindset, but rather a generic placebo text. Hence, we refer to it as "no-growth" rather than "fixed".

[12] Additional performance metrics available upon request.

4.4 Rating student letters

So far, we have classified student mindset based on their SMS responses, using different media to train our BERT model (hand-coded SMS responses, and the computer-based intervention scripts). Now, want to apply our BERT model to predict growth mindset in text written by students in different media, namely, student letters written as part of the tablet-based pilot conducted in 2019.

As part of this pilot, students were asked to write letters to future participants of the intervention. They were surveyed before and after the computer-based intervention ("baseline" and "end line" surveys) on a range of questions, including those related to growth mindset. As mentioned, these questions were just the Portuguese version of the English questions in Bettinger et al. (2018), whereby students had to rate the extent to which they agreed with each of the four statements in a Likert scale. We attribute a growth mindset to the letters written by students who weakly or strongly disagree with *all* the statements. All other letters are rated as 'no-growth' mindset, either because the students who wrote them have a fixed mindset (i.e., they weekly or strongly agree with all the questions), or because their mindset is undefined (i.e., they agree with some of the statements but disagree with others; following Claro et al., 2016).

Concretely, our goal is to predict growth mindset as revealed by the students in the 2019 pilot study survey, by classifying their letters written as part of this pilot using a BERT model trained on the SMS responses sent by the students in the study conducted in 2021. In addition to our main analysis, we employ a modified model that is trained on the combination of SMS by students and the intervention script.[13]

To evaluate the prediction performance across models in a concise way, we focus on accuracy and precision. The latter captures the ability to correctly identify growth mindset students, and the former, that to *also* identify students who are rated as having either a fixed or an undefined mindset – penalizing models that merely identify all students one way or the other. Table 5 lists accuracy and precision, calculated for each survey, treatment status, and BERT model. The reported accuracies range from 39 percent to 72 percent, with a notable pattern by survey and treatment status: while accuracy increases substantially in the treatment group between baseline and endline survey, it remains rather constant or at least increases to a smaller extent in the control group (the former being true for the model trained on SMS, and the latter being true for the model trained on SMS and script).

[13] We have also trained another BERT model using only the intervention scripts, with modest student letter prediction performance (results available upon request).

Table 5: Accuracy, by survey, treatment status, and BERT model

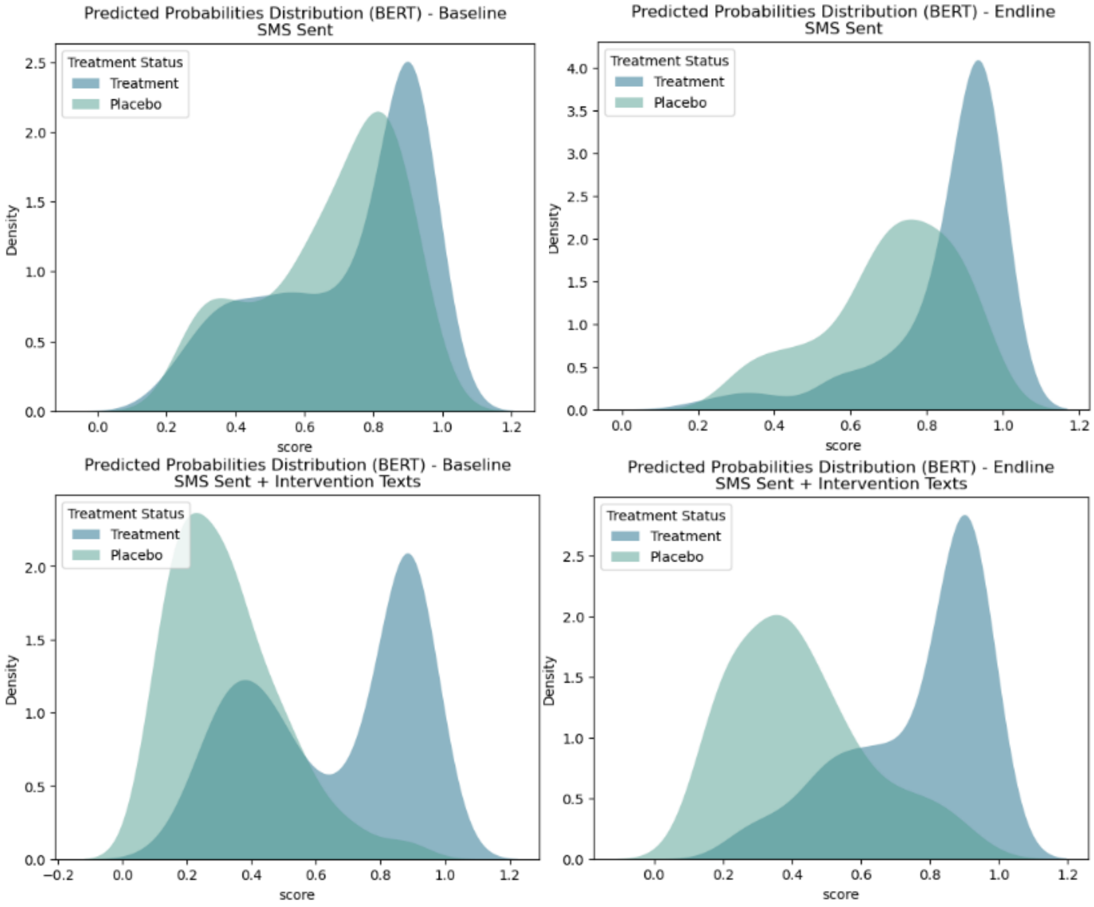
Survey	Treatment status	Metric	Large-sample models		
			SMS responses	Computer-based intervention script	SMS responses + Computer-based intervention script
Baseline	**Treatment**	**Accuracy**	0.628	0.697	0.541
		Precision	0.719	0.712	0.702
	Placebo	**Accuracy**	0.614	0.468	0.385
		Precision	0.698	0.71	0.769
End line	**Treatment**	**Accuracy**	0.715	0.732	0.672
		Precision	0.754	0.758	0.747
	Placebo	**Accuracy**	0.605	0.507	0.492
		Precision	0.633	0.591	0.681

Notes: Table presents accuracy and precision, calculated for each survey (baseline and end line) and treatment status (treatment and placebo), and underlying BERT model, trained on SMS sent by students; and on combination of SMS sent by students and the set of sentences of placebo (no-growth) / treatment (growth) intervention texts. Accuracy and precision based on comparison of BERT predictions, and the student replies when asked about their mindset (growth/no-growth). Predictions based on students' survey responses.

Furthermore, we can use these two BERT models to calculate the predicted probabilities that student letters belong to growth mindset, by survey and treatment group. At least for the model trained on the SMS responses only, Figure 8 showcases that predicted probabilities are clearly more divergent at end line. That is consistent with treatment effects on student mindsets, as discussed in Section 3.1.6 (there, estimated with survey responses as an outcome), except that, in this section, that conclusion is based solely on the content of student letters – a striking result.

All in all, we conclude that BERT-based prediction for student letters is also feasible, even if at lower levels of prediction accuracy as compared to the SMS prediction (Section 4.3).

Figure 8: Predicted probabilities, by survey, treatment status, and BERT model



Notes: Figure shows density of predicted probabilities of replies students wrote in the surveys (baseline/end line) belonging to a growth mindset, calculated for each survey (baseline and end line) and treatment status (treatment and placebo). The first row contains the distributions based on BERT model trained only on SMS sent by students, while the second row contains the distributions of a model also trained on the set of sentences of placebo (no-growth) / treatment (growth) intervention texts.

4.5 Rating growth mindset interventions

We further illustrate that our BERT toolkit can also be used to code how well SMS interventions sent out to students capture growth mindset. We do so by computing predicted probabilities that SMS are associated to growth mindset for 12 different SMS interventions evaluated in a companion study (Lichand et al., 2023a). In that paper, these interventions are grouped into two experiments. Experiment A focuses on the essential elements of the growth mindset message: (M1) the brain-muscle metaphor (‘your brain is like a muscle’, explored extensively in the computer-based intervention; see Appendix A), (M2) beliefs that returns to effort are high, (M3) beliefs that the costs of effort are low, (M4) future orientation, (M5) risk taking, and (M6) a sense of belonging in the school community. In turn, experiment B considers these very same dimensions, but embedded them in practical suggestions of how to get organized for studying as in-person classes gradually resumed over the first two school quarters of 2021.

We use two different BERT models to rate interventions. The first model is trained on the SMS that students sent (using the human annotations), as in Section 4.3. The second model is trained not only on the SMS that students sent, but also on the script of the computer-based intervention. Using both models separately, we predict the probabilities that a given SMS intervention captures growth mindset, and rank the 12 interventions accordingly.

Table 6 lists the 12 interventions, sorted by experiment (A or B), and, within experiment, by the predicted probabilities to capture growth mindset based on the first model. Overall, all interventions capture growth mindset rather well, with predicted probabilities ranging from 0.745 to 0.895 across models, experiments, and treatment arms. Although the predicted probabilities are not too different from each other, we can still produce a ranking of interventions. In experiment A, treatment arm M6 captures growth mindset the most, followed by treatment arm M3. This finding holds for the model trained only on SMS sent by students, and for the combined model of SMS and intervention scripts. In experiment B, treatment arm M2 captures growth mindset the most, followed by treatment arm M1. Here again, this result holds for both models. Although the full ranking of interventions is not identical for both models, they lead to similar conclusions overall.

Table 6: Predicting Growth Mindset Scores for SMS Interventions

Experiment	Treatment arm	SMS responses	SMS responses + Computer-based intervention script
A	**M6**	0.895 (0.121)	0.84 (0.149)
A	**M3**	0.892 (0.121)	0.838 (0.152)
A	**M1**	0.889 (0.12)	0.825 (0.149)
A	**M5**	0.888 (0.12)	0.812 (0.147)
A	**M2**	0.884 (0.088)	0.85 (0.131)
A	**M4**	0.881 (0.12)	0.805 (0.143)
B	**M2**	0.883 (0.081)	0.85 (0.106)
B	**M1**	0.881 (0.095)	0.836 (0.127)
B	**M4**	0.871 (0.081)	0.745 (0.149)
B	**M5**	0.87 (0.104)	0.793 (0.147)
B	**M6**	0.858 (0.087)	0.811 (0.108)
B	**M3**	0.851 (0.105)	0.767 (0.164)

Notes: Table presents the mean and standard deviation of predicted probabilities that given SMS intervention captures growth mindset, by experiment (A and B), treatment arm (M1 – M6), and model (BERT model trained on SMS sent by students, and BERT model trained on SMS sent by students and scripts of computer-based intervention).

A relevant pattern is that, in most cases, the fidelity scores for experiment A are higher than for experiment B. This might help rationalize the findings in Lichand et al. (2023a), which documents that only experiment A significantly increased the prevalence of a growth mindset among targeted students and improved educational outcomes as a result.

5. Discussion

This paper develops a toolkit for measuring student mindsets in low-resource settings. Our findings from a large-scale survey based on this toolkit match those of face-to-face surveys collected in Chile, validating our survey methodology.

The paper also trains a natural language machine learning model to automatically classify students as fixed or growth mindset based on their SMS responses to our question prompts. The model displays high precision and accuracy relative to hand-coding, showing promise for applications with much larger samples or higher response rates (which are not amenable to manual coding of student responses).

Such model was then successfully applied to rate student letters in a previous pilot. This application showcases the power of the model, since it captures treatment effects of a previous intervention, conducted during a pilot study 1.5 years before the data on which the model was trained. Last, the model was also applied to rate growth mindset SMS interventions. Results once again showcase the model's potential, as it attributes higher fidelity precisely to the set of growth mindset interventions that were most effective in changing student mindsets and improving their educational outcomes in a companion study.

This toolkit naturally has several limitations. We study a specific application, based on SMS surveys. While we expect it to be widely applicable, since SMS works in any functioning phone (even without connectivity or even airtime), electricity constraints or other issues might prevent the application of phone surveys in specific settings, e.g., among displaced populations. Focus group participants also noticed that, even when SMS is available, it might not be the most effective tool to engage students. In fact, we have experienced substantially low response rates, and even more complex challenges in sustaining participation over time, leading to a selected sample relative to the universe of students with phones on record.

Having said that, our results provide evidence that the classifier algorithm trained on the SMS data is applicable even more broadly: it was able to accurately predict student mindsets when applied to text based on letters written by students using a completely different media. Taken together, the different analyses provide a blueprint of how our SMS toolkit could be applied in practice, to SMS responses and beyond. Multimedia applications, that combine SMS with other media to broaden the pool of respondents, might be particularly promising.

Another important caveat is that while results are encouraging about the possibility of applying our classifiers to other use cases, the sensitivity of the classifier to using the 'voices' of the target population itself (since the model trained on students' SMS responses dominates that merely trained on the computer-based intervention script, even after the latter was adapted to the local context) suggests that, if given the opportunity, researchers should train their own BERT models in their study setting, taking advantage of the step-by-step BERT architecture discussed in Appendix D.

Our BERT model performs well in automatically classifying student mindsets based on their SMS responses. However, the classification accuracy may depend on the medium used to train the model, student subgroups and a range of coding decisions, highlighting the importance of a careful application of classification technologies. If calibrated successfully, NLP-based technologies like our SMS toolkit can be used to predict features in text other than SMS written by students too, as exemplified by the mindset prediction of the student letters and the intervention SMS sent to students.

At a general level, our results demonstrate that it is possible to track student mindsets even in low-resource settings, as long as context-appropriate survey instruments and technologies are used. Further research is needed to shed light on the appropriateness of different instruments and technologies to different study populations.

ACKNOWLEDGEMENTS

We acknowledge the invaluable inputs of the Rio de Janeiro State and São Paulo State Secretariats of Education staff members, of the participants of the numerous focus groups conducted between 2019 and 2023, and of the respondents of the SMS surveys conducted as part of this study. All remaining errors are ours.

RESEARCH TRANSPARENCY AND ETHICAL APPROVAL

Replication package available upon request. Survey instruments preregistered as part of trials 6436 and 7152 at the AEA Social Science Registry and approved by the University of Zurich Institutional Review Board.

FUNDING AND CONFLICT OF INTEREST

Research generously funded by the Stanford University Lemann Center for Brazilian Education and Entrepreneurship. Guilherme is a co-founder and chairman at Movva, the implementing partner of the SMS surveys used in this study.

REFLEXIVITY STATEMENT

Based on the framework in Jacobson and Mustafa (2019), we reflected on the positionality of each of the authors relative to the subject matter of this study, and relative to our study population. GL, CD and AT are all Brazilian, grew up in the country, and have been actively engaged with research in education in Brazil since many years now. Nevertheless, they all declare themselves as white and upper middle-class, having studied in private schools, and hence removed from the public-school setting that circumscribes our study participants. JG, in turn, is Colombian. He declares himself as Latino and middle class. His experience in Colombia is similar to that of the Brazilian co-authors – engaged with the issues of public education through his research, although not directly as a student in that system. For all the Latin American authors, being born and raised in the region gives them perspective about the subject matter of this study, as inequalities in educational access, quality and subsequent returns are prominently feature in everyday life in these countries. Last, EA, EB and DY are American, and BA is German. All of them declare themselves as white and (upper) middle class. While they did not directly or indirectly experience the educational journeys of our study population, they have all been long involved in research on international and comparative education, in particular when it comes to documenting and addressing educational inequalities. The lack of first-hand experience of the research team with students' journeys in the Brazilian public education system led us to rely extensively on local partners, particularly the Rio and São Paulo Education Secretariat staff members, and on the several focus groups discussions both prior to running the study, in our consecutive pilots, and after obtaining results, in order to validate our methodology and results with our study population.

REFERENCES

Brossard, M., Carnelli, M., Chaudron, S., Di-Gioia, R. Dreesen, T., Kardefelt-Winther, D., Little, C. and Yameogo, J. (2021). Digital Learning for Every Child: Closing the Gap. https://www.unicef.org/media/113896/file/Digital%20. Accessed: 05/10/2022.

Claro, S., Paunesku, D., & Dweck, C. S. (2016). Growth Mindset Tempers the Effects of Poverty on Academic Achievement. *Proceedings of the National Academy of Sciences*, 113, 8664-8668. https://doi.org/10.1073/pnas.1608207113.

Devlin, J., Chang, M-W., Lee, K. and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint,* arXiv:1810.04805.

Dweck, C. S. (2006). *Mindset: The new psychology of success.* Random House.

Ganimian, A. J. (2020). Growth-Mindset Interventions at Scale: Experimental Evidence From Argentina. *Educational Evaluation and Policy Analysis*, 42(3), 417–438. https://doi.org/10.3102/0162373720938041.

Hecht, C., Dweck, C., Murphy, M., Kroeper, K. and Yeager, D. (2023). Efficiently exploring the causal role of contextual moderators in behavioral science. PNAS, 120(1), e2216315120.

Hecht, C., Murphy, M., Dweck, C., Bryan, C., Trzesniewski, K., Medrano, F., Giani, M., Mhatre, P. and Yeager, D. (2023). Shifting Classroom Cultures to Address Educational Disparities, *mimeo*.

Jacobson, D. and Mustafa, N. (2019). Social Identity Map: A Reflexivity Tool for Practicing Explicit Positionality in Critical Qualitative Research. International Journal of Qualitative methods, 18, 1-12. DOI: 10.1177/1609406919870075

Lichand, G., Cunha, N., Madeira, R. and Bettinger, E., (2022). When the Effects of Informational Interventions Are Driven by Salience: Evidence from School Parents in Brazil. SSRN Working Paper, 3644124, January 2022.

Lichand, G., Bettinger, E., Trindade, A. Belchior, C., Rege, M., and Yeager, D (2023a). What Is It About the Growth Mindset Intervention? Experimental Evidence from an SMS Intervention in Brazil, *mimeo*.

Lichand, G., Ash, E., Arold, B., Gudino, J., Belchior, C., Rege, M., Trindade, A. Bettinger, E., and Yeager, D (2023b). The Opportunities and Challenges of ML-based Targeting for Educational Interventions: Evidence from Brazil in the Aftermath of the Pandemic, *mimeo*.

PWC (2022). O Abismo Digital no Brasil. https://www.pwc.com.br/pt/estudos/preocupacoes-ceos/mais-temas/2022/o-abismo-digital-no-brasil.html. Accessed: 05/10/2023.

Rege, M., Hanselman, P., Solli, I., Dweck, C., Ludvigsen, S., Bettinger, E., Crosnoe, R., Muller, C., Walton, G., Duckworth, A., Yeager, D. (2020). How can we inspire nations of learners? Investigating growth mindset and challenge-seeking in two countries. *American Psychologist*. 76. 10.1037/amp0000647.

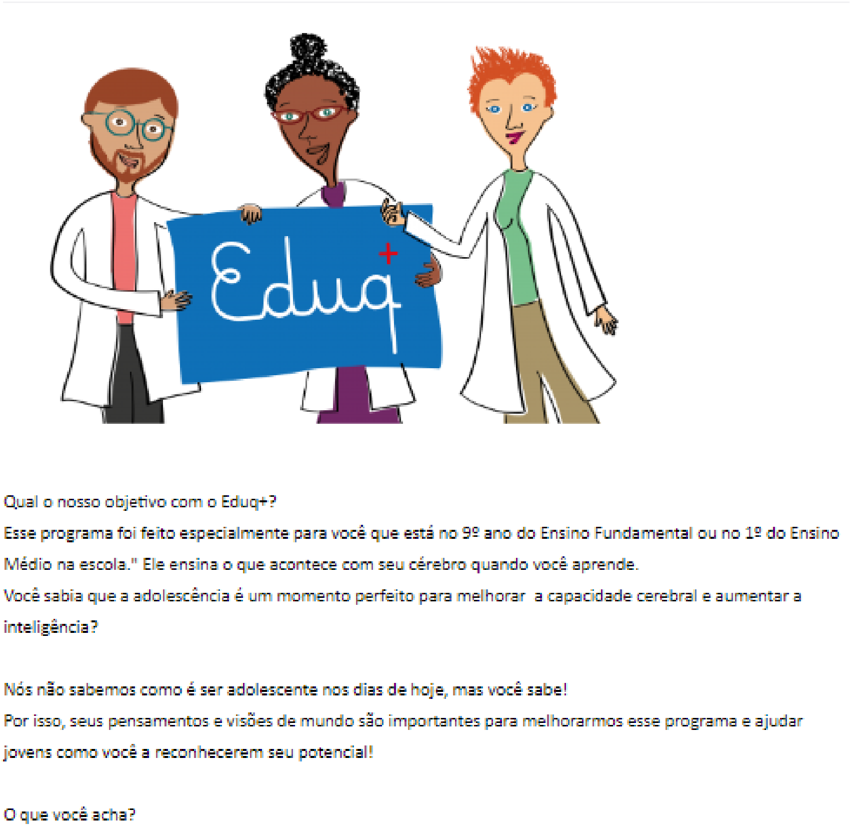
TIC (2020). Pesquisa Educação 2020. *Available at*: https://cetic.br/pt/pesquisa/educacao/. Accessed: 11/14/2022.

Yeager, D.S., Hanselman, P., Walton, G.M. *et al* (2019). A national experiment reveals where a growth mindset improves achievement. *Nature,* 573, 364–369. https://doi.org/10.1038/s41586-019-1466-y.

Yeager, D.S., Bryan, C.J., Gross, J.J. *et al* (2022). A synergistic mindsets intervention protects adolescents from stress. *Nature,* 607, 512–520. https://doi.org/10.1038/s41586-022-04907-7.

Appendix A – Screenshots from the adapted computer-based intervention

Figure A1. Introductory screen



Qual o nosso objetivo com o Eduq+?

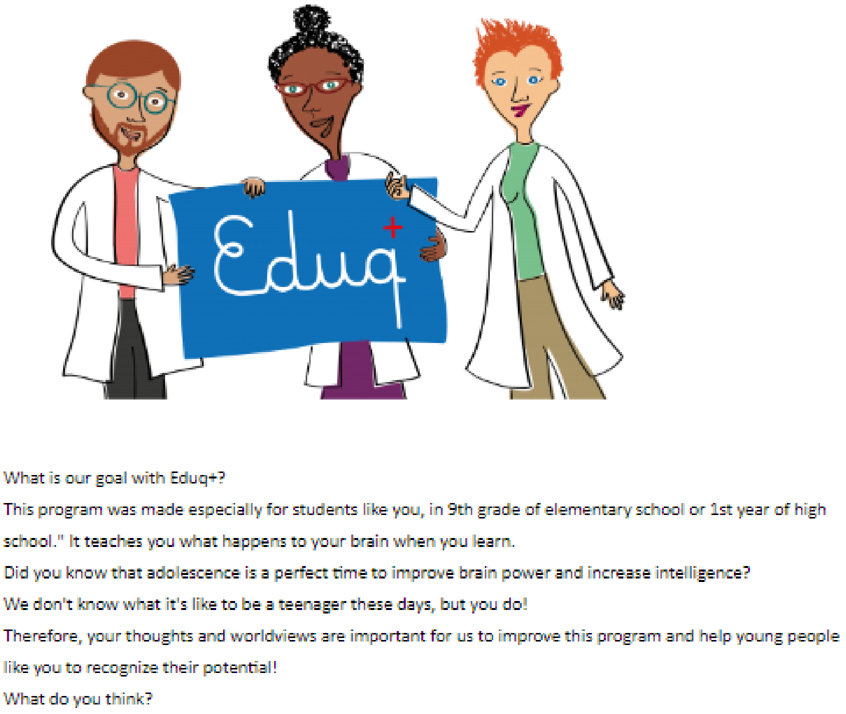
Esse programa foi feito especialmente para você que está no 9º ano do Ensino Fundamental ou no 1º do Ensino Médio na escola." Ele ensina o que acontece com seu cérebro quando você aprende.

Você sabia que a adolescência é um momento perfeito para melhorar a capacidade cerebral e aumentar a inteligência?

Nós não sabemos como é ser adolescente nos dias de hoje, mas você sabe!

Por isso, seus pensamentos e visões de mundo são importantes para melhorarmos esse programa e ajudar jovens como você a reconhecerem seu potencial!

O que você acha?



What is our goal with Eduq+?

This program was made especially for students like you, in 9th grade of elementary school or 1st year of high school." It teaches you what happens to your brain when you learn.

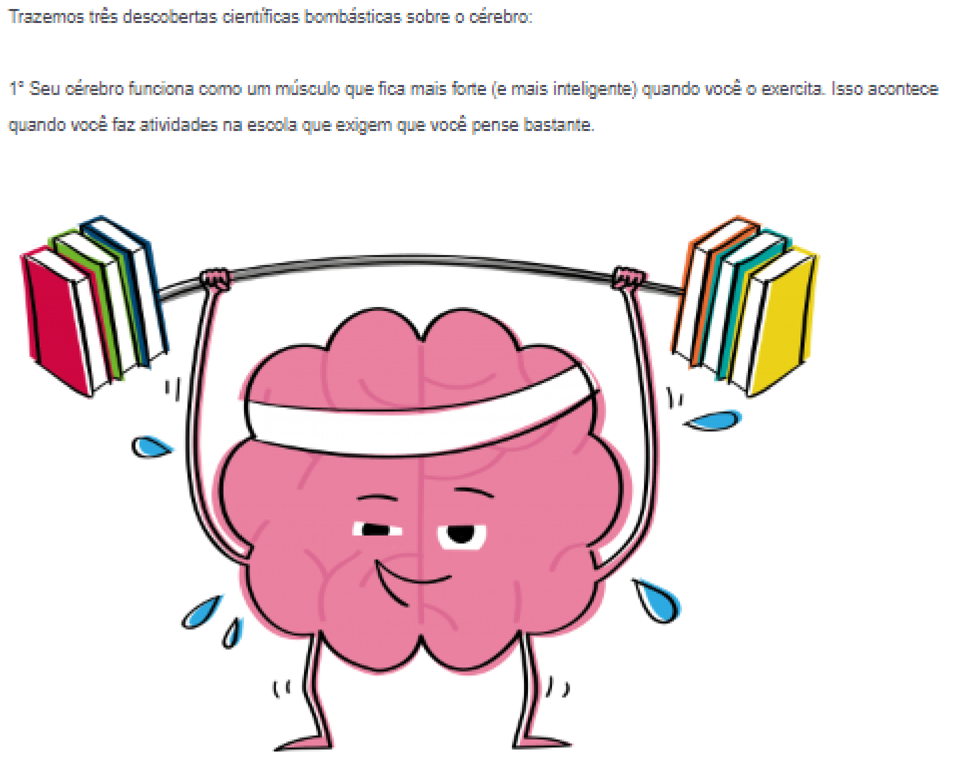
Did you know that adolescence is a perfect time to improve brain power and increase intelligence?

We don't know what it's like to be a teenager these days, but you do!

Therefore, your thoughts and worldviews are important for us to improve this program and help young people like you to recognize their potential!

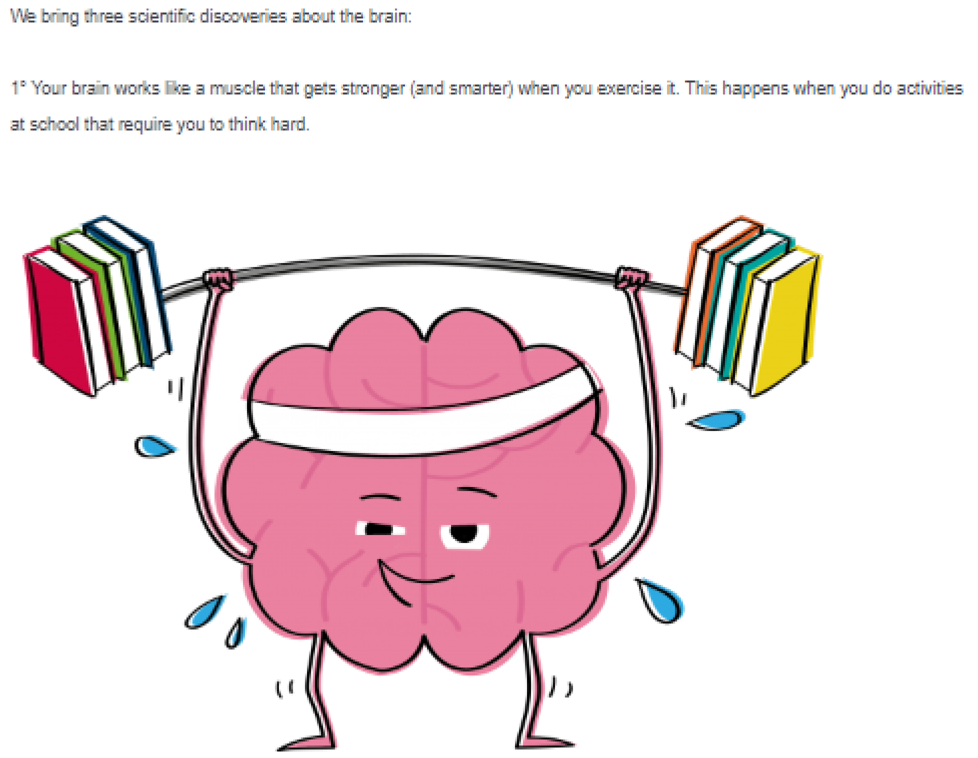
What do you think?

Figure A2. Treatment version



Trazemos três descobertas científicas bombásticas sobre o cérebro:

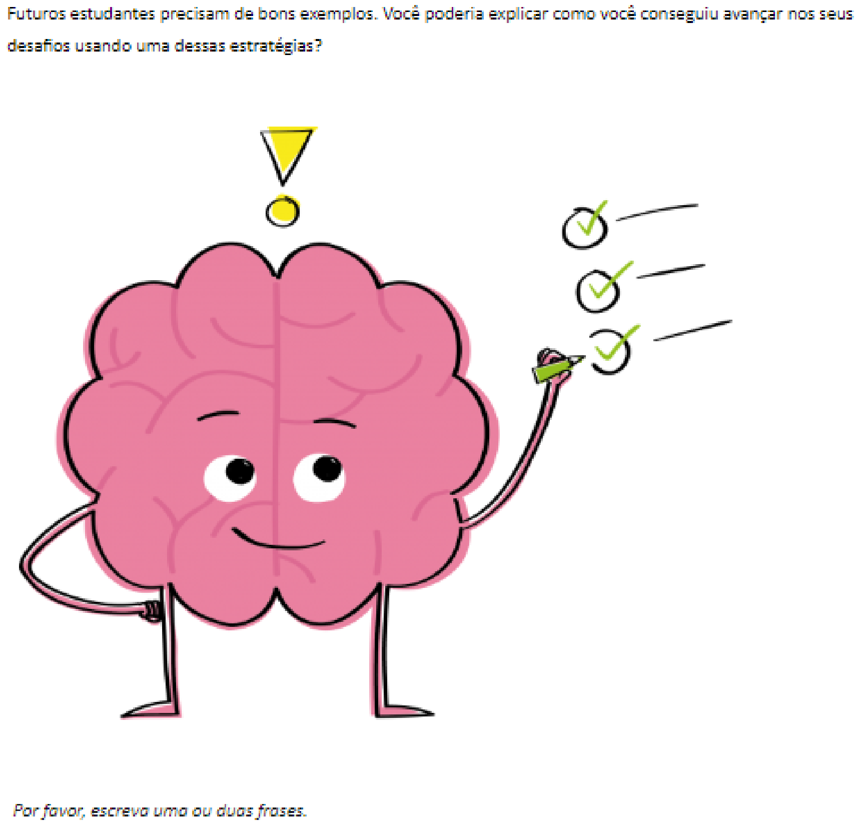
1° Seu cérebro funciona como um músculo que fica mais forte (e mais inteligente) quando você o exercita. Isso acontece quando você faz atividades na escola que exigem que você pense bastante.



We bring three scientific discoveries about the brain:

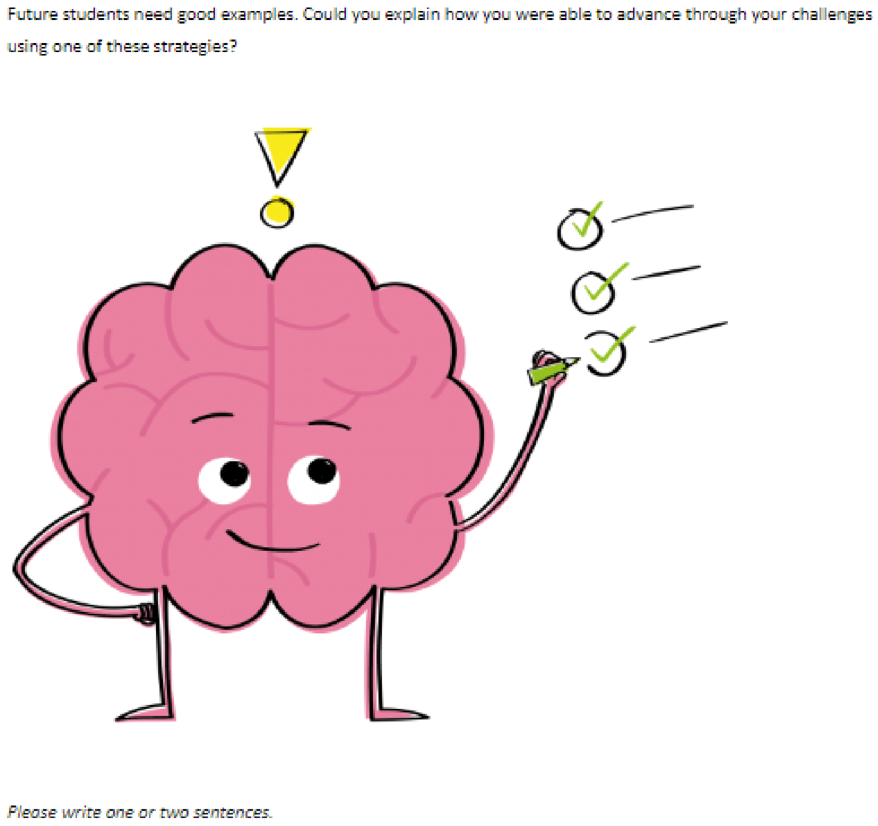
1° Your brain works like a muscle that gets stronger (and smarter) when you exercise it. This happens when you do activities at school that require you to think hard.

Figure A3. Prompts for student letters



Futuros estudantes precisam de bons exemplos. Você poderia explicar como você conseguiu avançar nos seus desafios usando uma dessas estratégias?

Por favor, escreva uma ou duas frases.

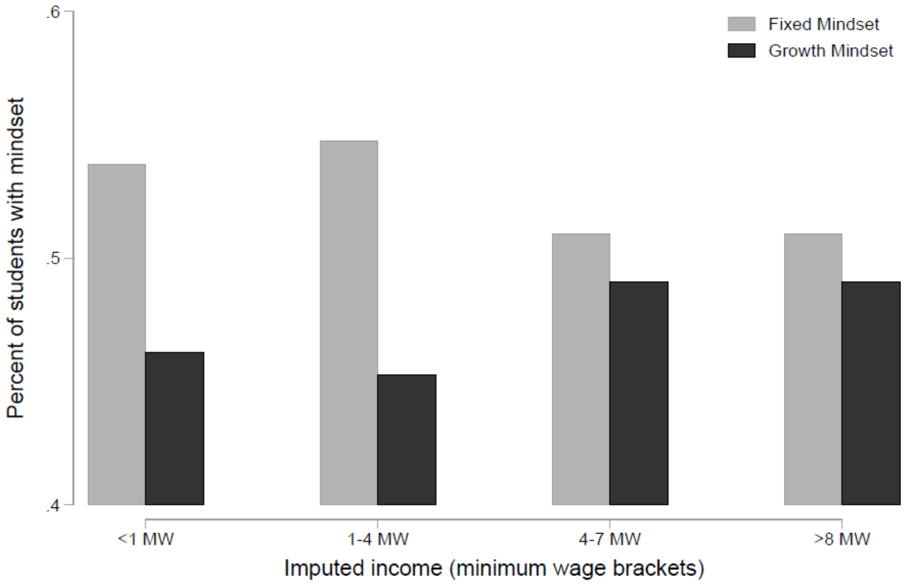


Future students need good examples. Could you explain how you were able to advance through your challenges using one of these strategies?

Please write one or two sentences.

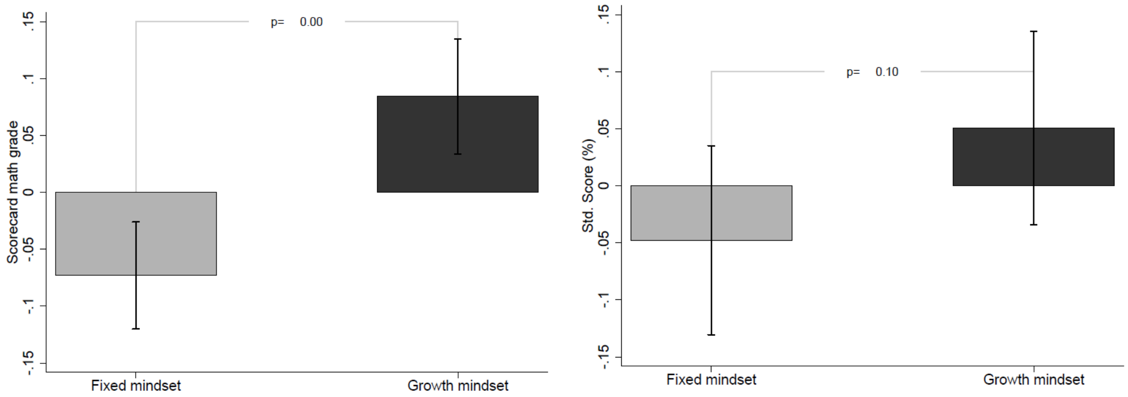
Appendix B – Additional results

Figure B1. Robustness to re-weighting observations to match universe of students



Notes:

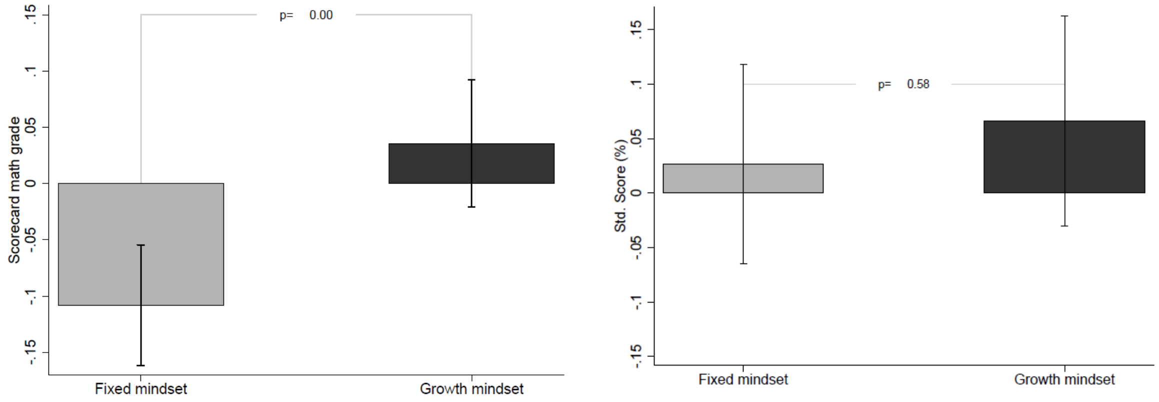
Figure B2. Association between test scores and student mindsets



(a) Report card math grade (b) Math std. test scores

Notes:

Figure B3. Robustness to re-weighting observations to match universe of students

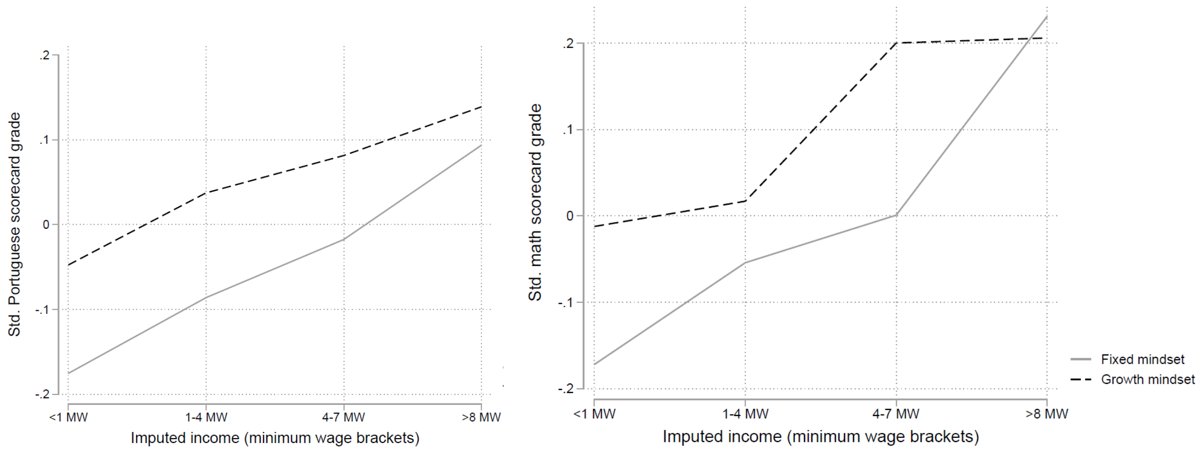


(b) Report card math grade

(b) Math std. test scores

Notes:

Figure B4. Robustness to re-weighting observations to match universe of students



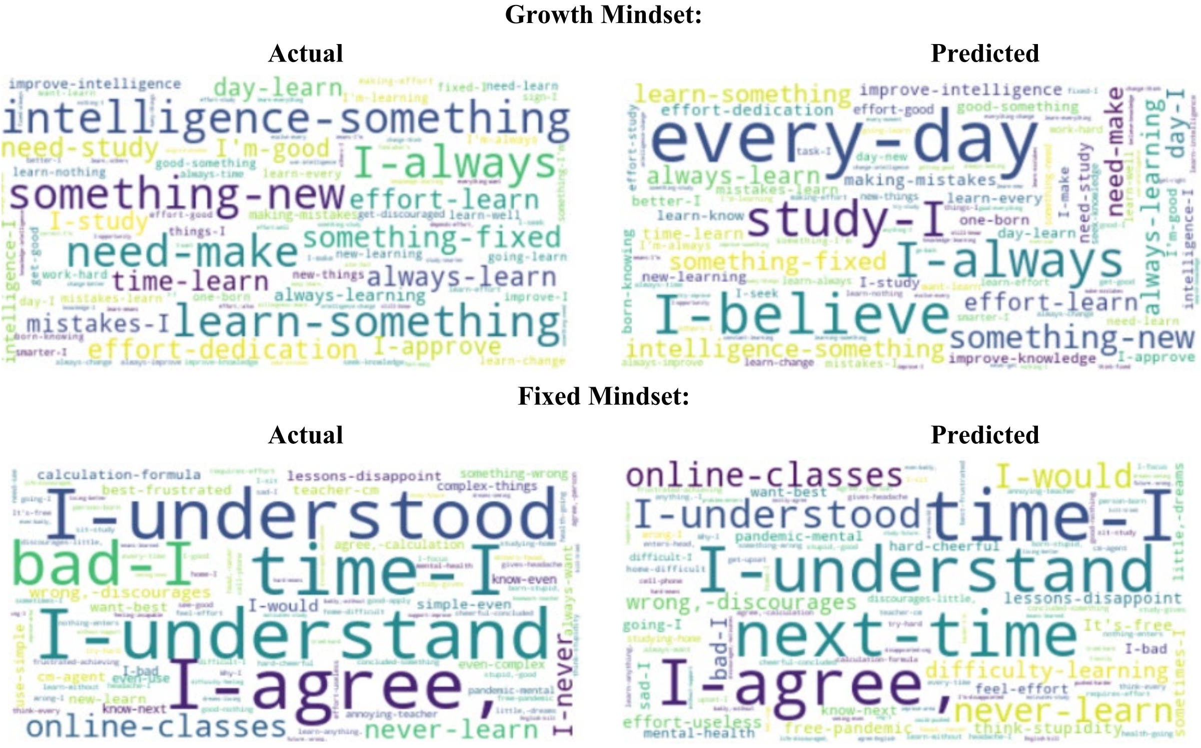
(a) Report card Port. grade

(b) Report card math grade

Notes:

Appendix C – Additional results for machine learning classifier

Figure B1. Bigrams: growth mindset (upper graphs) and fixed mindset (lower graphs)



Growth Mindset:

Actual

Predicted

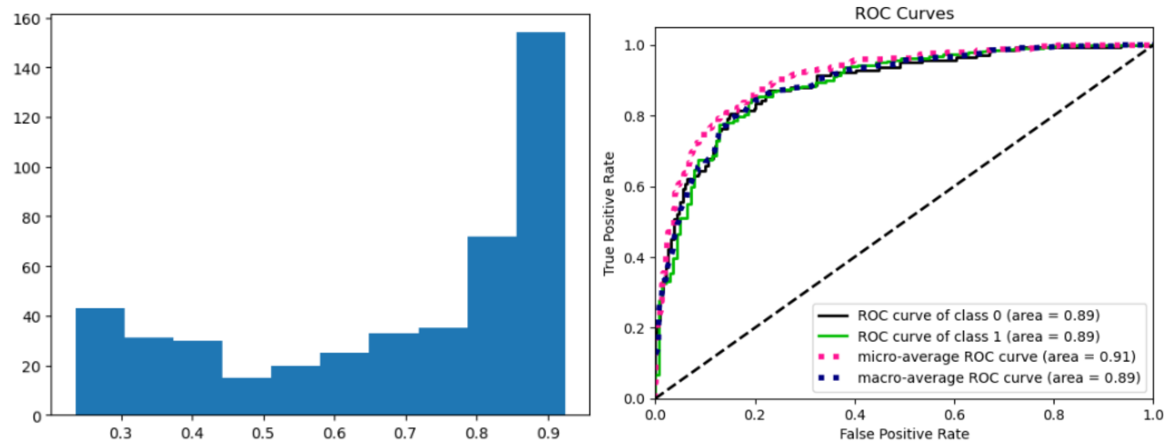
Fixed Mindset:

Actual

Predicted

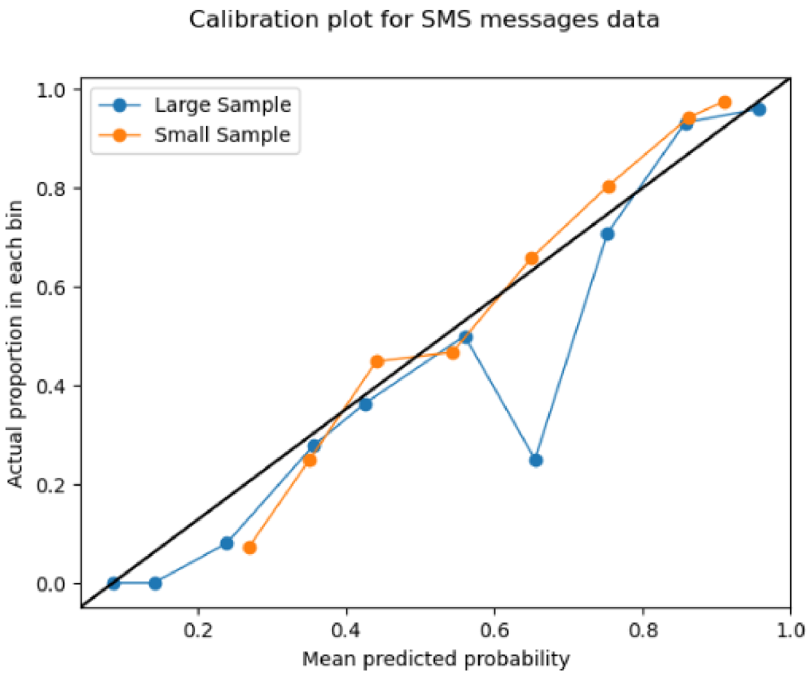
Notes: Figure shows word clouds of bigrams, by mindset (growth vs. fixed) and model ("actual" human annotator vs. "predicted" BERT) based on relative frequencies. The larger the font size, the more important is a bigram for a given mindset in relative terms.

Figure C2. Probability distribution (LHS) and ROC curves (RHS), small sample



Notes: Figure on the LHS depicts histogram of the estimated probability of being a growth-mindset student, according to BERT model. Figure on the RHS depicts ROC (Receiver Operating Characteristic) curve, using different thresholds for the probability vector to measure goodness of fit. The ROC curve for class 0.0 represents the performance of the model in correctly classifying instances of growth mindset, while the ROC curve for class 1.0 represents the performance of the model in correctly classifying instances with label fixed mindset. The micro-average ROC curve aggregates the TPR and FPR across growth mindset and fixed mindset, and produces a single ROC curve. The macro-average ROC curve computes the TPR and FPR for growth mindset and fixed mindset separately and then takes the average across the two classes.

Figure C3. Calibration plot



Notes: Figure depicts calibration plot for two BERT models, trained on a large sample (blue) and small sample (orange).

Table C1. Confusion matrix, small sample

Actual\Predicted	Fixed	Growth
Fixed	16	19
Growth	6	82

Notes: Table presents the confusion matrix of actual and predicted growth/fixed mindset SMS. Actual classification performed by human annotators, predicted classification performed by BERT algorithm. Model training was implemented based on 70/30 random split for SMS messages larger than 30 characters.

Appendix D – BERT architecture

To implement the BERT model, we conduct the following three steps:

1. We start by using the BERTimbau[14] model to calculate a numerical vector of 768 dimensions (embeddings) for each token of each message/paragraph. We use its base version due to computational efficiency. The model yields a sequence of vectors for each message. The accuracy of these numerical representations is relatively high as they were calculated using brWaC corpus, which contains 2.68 billions of tokens from 3.53 millions of documents. It is the largest Portuguese corpus to date (Souza et al, 2019).

2. Using each sequence of embeddings, we train a BERT model (Devlin et al, 2018). Intuitively, our BERT model transforms each embedding in each sequence into a weighted average of the embeddings of the sequence (self-attention function), where weights are estimated by a neural network that classifies the sequences of tokens into growth or non-growth/fixed mindset. These weights, based on cosine distances, ultimately allow the algorithm to distinguish mindsets.

3. Before implementing the classification tasks, we fine-tune our model. Following standard practices, we explore which combinations of values of learning rates (2e-5, 3e-5, 5e-5), number of epochs (1,2,3), weight decay (0.001,0), and batch size (16,32) give the highest accuracy. Setting the number of warmup steps to 0, we find that the optimal combination sets the learning rate to 2e-5, the number of epochs to 3, the weight decay to 0, and the batch size to 16 for all models. We perform training/development splits of 70/30 for different combinations of datasets.

[14] For more information, see: https://huggingface.co/neuralmind/bert-base-portuguese-cased