Article

Do Written Responses to Open-Ended Questions on Fourth-Grade Formative Assessments in Mathematics Help Predict Scores on End-of-Year Standardized Tests?

Felipe Urrutia [1,†,‡], **Roberto Araya** [1,‡]

[1] Affiliation 1; e-mail@e-mail.com
[2] Affiliation 2; e-mail@e-mail.com
* Correspondence: e-mail@e-mail.com; Tel.: (optional; include country code; if there are multiple corresponding authors, add author initials) +xx-xxxx-xxx-xxxx (F.L.)
† Current address: Affiliation 3
‡ These authors contributed equally to this work.

Abstract: Predicting long-term student learning is a critical task for teachers and for educational data mining. However, most of the models do not consider two typical situations in real-life classrooms. The first is that teachers develop their own questions for formative a ssessment. Therefore, there are a huge number of possible questions, each of which is answered by only a few students. Second, formative assessment often involved open-ended questions that students answer in writing. These types of questions in formative assessment are highly valuable. However, analyzing the responses automatically can be a complex process. In this paper, we address these two challenges. We analyzed 621,575 answers to closed-ended questions and 16,618 answers to open-ended questions by 464 fourth-graders from 24 low-SES schools. We constructed a classifier to detect incoherent responses to open-ended mathematics questions. We then used it in a model to predict scores on an end-of-year national standardized test. We found that despite answering 36.4 times fewer open-ended questions than closed questions, including features of the students' open responses in our model improved our prediction of their end-of-year test scores. To the best of our knowledge, this is the first time that a predictor of end-of-year test scores has been improved by using automatically detected features of answers to open-ended questions on formative assessments.

Keywords: Computational linguistics; elementary mathematics; formative assessments; student models

1. Introduction

One of the most recommended teaching and learning strategies is formative assessment [27], [30]. These are quizzes or low/no-stakes assessments. Although the main audience of formative assessments are students, they are also critical for teachers [29]. They provide timely information to the teacher on the status of the learning process and an estimate of the state of knowledge attained by each student. However, it is important to distinguish between the knowledge attained immediately after learning activities and the definitive knowledge that will be revealed by students in the long term. There is a big difference between what the student demonstrates during or shortly after attending a lesson, and what she reveals months or years later. Estimating long-term learning is a major challenge for the teacher. This is because there are many examples of strategies that generate learning when measured immediately after the intervention, but demonstrate a rapid decline in the long run. There are also interventions where the opposite is true [33]. In these alternative interventions, students are exposed to a series of pre-designed and desirable difficulties [26]. These interventions require more effort from the student and lead to slower progress. Subsequently, students only manage a weak performance in the short term. However, they eventually produce a strong performance in the long term. Long-term learning in this kind of intervention is therefore better than in the first kind of intervention,

 doi:10.20944/preprints202204.0002.v1

described previously. This reversal phenomenon is difficult to swallow. According to [33], what we can observe and measure during teaching is performance, which is often an unreliable index of long-term learning. [33] make a critical distinction between performance, as measured during acquisition, and learning, as measured by long-term retention or transfer capacity. This is an unintuitive phenomenon, where fast progress generates an illusion of mastery in the students [26]. This illusion also hinders the teacher and makes it difficult for her to make a good prediction of each of her students' long-term learning.

In this paper, we use a database of questions and answers taken from ConectaIdeas [21], [38], [25]. This is an online platform where students can answer closed- and open-ended questions. Teachers on ConectaIdeas can develop their own questions, designing them from scratch or taking them from existing material, or select them from a library of questions designed previously by other teachers. The teachers then use those questions to build their own formative assessments composed of sets of 20 to 30 questions. Students either answer them in laboratory sessions held once or twice a week, or at home.

In principle, open-ended questions allow teachers to visualize their students' reasoning. This is particularly true if the question asks for explanations. These are questions that require more effort from students, but which are posed much less frequently than closed questions. Moreover, these are teacher-adapted or teacher-designed questions, thus leading to a wide range of question types. It is therefore more challenging to estimate their long-term learning than when working with questions taken from a carefully-crafted and small list of closed questions. In this paper, we examine the responses of fourth-grade students from at-risk schools. The answers are very short. The average number of words in the responses is 8-9 words. This average increases as the school year progresses [23],[22], [25]. On the other hand, in previous RCT studies using the ConectaIdeas platform with fourth graders, we have found that the length of the responses to open-ended questions has a significant and positive effect on end-of-year learning in math [25].

Research question: To what extent do students' short, written answers to teacher-designed, open-ended questions in weekly formative tests help improve predictions of performance on end-of-the-year national multiple-choice standardized assessments?

2. Related works

[40] developed linear predictors of student scores on end-of-year state tests using dynamic testing metrics developed from monthly log data taken from an intelligent tutoring system. They analyzed the data logs for 362 students, although only 105 students had complete data in each of the months. They found that logs from an online tutoring system provide better end-of-year predictions than using paper and pencil benchmark tests. They found that the adjusted R^2 is 0.637. However, all students attempted to solve a similar set of items. One of the challenges is that teachers prefer having the flexibility to adapt the exercises to their own experience. The authors therefore did not consider the situation in which teachers select or design their own set of exercises. Moreover, the study did not include answers to open-ended questions.

From a sample of 23,000 learners in Grades 6, 7, and 8 over three academic years, [40] analyzes the relative contribution of different types of learner data to statistical models that predict scores on summative assessments. They use six different categories of statistical models to predict summative scores. One of the best model categories turned out to be Stepwise Linear Regression (SLR). In the best year, it achieved an R2 of 0.734. However, this study does not consider variables associated with answers to open-ended questions. Apparently, the platform used does not include this type of question.

[28] analyzed 10 learning techniques and considered four categories of variables: learning conditions, student characteristics, materials, and criterion tasks. The techniques include elaborative interrogation, self-explanation, summarization, highlighting (or underlining), the keyword mnemonic, imagery use for text learning, rereading, practice testing, distributed practice, and interleaved practice. Two of these techniques are related to explaining. Elaborative interrogation is defined as generating an explanation for why an

 doi:10.20944/preprints202204.0002.v1

explicitly stated fact or concept is true. Self-explanation is defined as explaining how new information is related to known information or explaining steps taken. They found that self-explanation has a moderate effect and that a major strength is that its effects have been shown across a wide range of content materials and tasks. However, in their review, they do not study the effect of written explanations.

[32] conducted a Meta-Analysis of the use of Self-Explanation to Improve Mathematics Learning. These are self-explanations generated by the learner and without the intent of teaching to someone else. They reviewed 26 published studies that contrasted prompts to self-explain with a control condition involving unprompted explanations. They found a statistically significant effect (p-values less than .05) when testing is immediate, but not after there is a delay.

[34] propose a novel deep learning framework to effectively integrate both question explanations and student responses for predicting students' current learning outcomes. They use the responses in two exams for training and try to predict responses in a third exam, taken after the first two. Each exam has a total of 46 questions. The first two exams were taken by 5,675 and 6,895 students, respectively, while the third was taken by 26,330 students. In this sense, there was a small and fixed number of questions that were carefully designed by experts, as well as a large number of students who answered them. This allows for patterns to be discovered using big data algorithms. In our paper, on the other hand, the questions are practically not repeated between different courses, since they are designed by the teacher. Although the teacher can copy them or adapt them from other teachers' questions, they are very rarely copied verbatim. This condition is much more frequent and naturally-occurring in classrooms, where the teacher decides on the spot to pose questions that they deem appropriate in that specific moment.

[39] studied the effect of writing on the mathematical problem-solving skills of 96 middle school students who participated in a 6-week afterschool program. The study compared the impact with a randomly assigned control group, who prepared for a high-stakes test involving mathematical problem-solving. The authors found that students from the experimental group were more likely to generate and apply better problem-solving skills than the control group. Indeed, they outperformed the control group on a test of cognitive complexity and problem generation. It can therefore be inferred that there is some empirical evidence of the effect of writing on mathematics. However, for the control group it is not clear whether they did practice tests using closed questions. In addition, the materials and the post-tests were designed by the research team. They were not end-of-year, state standardized summative tests.

[31] compare the effects on learning of multiple-choice and short-answer quizzes on. According to the authors there is empirical evidence that more activated or engaged processes lead to improved retention on final summative tests. Multiple-choice tests rely more on recognition than recall, while laboratory studies have found that short answers lead to more long-term learning. However, their comparison study in a classroom setting found no clear advantages of short answers. This is a puzzling finding that contradicts the empirical evidence obtained in laboratory studies. In any case, short answers in this study only involved single-word responses. Students were not required to provide a sentence explaining a result. Instead, students needed to complete a sentence by suggesting the missing word.

3. Materials and Methods

Our data comes from a virtual platform called ConectaIdeas. This platform has a series of mathematics exercises for elementary school students. Each student answers two types of questions during sessions at specific times of the year. The questions are created by teachers and can be either closed-ended (e.g., multiple choice questions) or open-ended (e.g., essay questions). Throughout the school year, students answer about 30 closed-ended questions and one or two open-ended questions during each session. The sessions are 90 minutes long, with two sessions held per week. Unlike the answers to closed-ended questions, the

 doi:10.20944/preprints202204.0002.v1

data for the answers to the open-ended questions is unstructured and provided in the form of written text. This is a particularly challenging feature.

Below, we outline the materials and methods used for two problems: detecting the coherence of answers to open-ended questions, and predicting the scores on an end-of-year national standardized test.

3.1. Detecting the coherence of answers to open-ended questions

3.1.1. Data

For each session on the platform, we collected the open-ended questions set by the teachers and the written responses provided by the students. The data set consists of a compilation of sessions from 2017 and 2019. For both years, the question-answer format is the same but only the data for 2019 were labeled. The objective is to fit a prediction model using the 2019 data in order to predict the 2017 data.

The questions are created by teachers and belong to the five strands of the national mathematics curriculum. Each one has a statement that introduces the problem and concludes with a key question (e.g. how much is it?). These questions are very diverse in morphology (i.e. how to ask) and learning objectives (i.e. what to ask). In some cases, only decontextualized quantities are introduced. In others cases, situations and characters are also introduced. Likewise, some of them conclude with questions that require justification or an explanation, while others do not. With this, six types of questions have been typified, defined as follows:

1. *Calculate without explaining*. They consist of questions that ask to calculate a quantity but do not ask for an explanation or justification of the answer.
2. *Calculate with explaining*. Unlike the previous type of question, this type of question asks to explain, justify or demonstrate how or why such a result was obtained.
3. *Choice and/or affirmation*. This type of question introduces characters and statements, and can take one of two forms. In the first, more than one character is introduced, with each giving a statement. The student then has to choose who is right. In the second, a character is presented with a statement and the student has to state whether or not they are right. In addition, all of these questions ask for a justification of the answer given.
4. *Compare quantities*. Most of these questions mention two quantities and ask to name which or why these quantities are equal or why one is greater than the other.
5. *Procedure and content knowledge*. These consist of two types of question. The first seek to exemplify using an invented problem. The others are content questions (e.g. What are the axes of symmetry?)
0. *Others*. This type of question is used to indicate any questions that do not belong to any of the other categories.

All 2019 questions were tagged by a member of the research team based on the six types of question. We reviewed these labels and checked that they were well classified according to their definition. Examples of each type of question can be seen in Table 1.

 doi:10.20944/preprints202204.0002.v1

Table 1. Example for each type of question.

Type of question	Question
1	If Mariela has 30 flowers and her friend Juanita takes 10 flowers from her, How many flowers does Mariela now have?
2	Lorena has a $5,000 bill and wants to exchange it for $500 coins. How many $500 coins would Lorena have? Explain your answer in your own words.
3	Luisa has ninety-eight sheets, her brother says that this number is written as 908. Is Luisa's brother correct?
4	Martina and Juan each bought a pizza of equal size. Martina ate 3/8 of her pizza and Juan ate 3/4 of his. Who ate more pizza? Explain in your own words how you arrived at the result.
5	What is an axis of symmetry? Explain in your own words and give me an example.
0	Daniel buys 856 cakes for his business. Write this number in WORDS.

[1] Originally in Spanish

The questions were answered by elementary school students, and the type of answer varied depending on the type of question. If a quantity is asked for with/without explanation, answers are expected with numerical representations (e.g. integers, decimals, fractions, etc.) and with/without arguments, as appropriate. In other cases, if asked about a character and their statement, answers including at least the name of a character and a key affirmative word (e.g., yes or no) are expected. In some cases, what the students answer does not make sense with what they are being asked (e.g., they are asked to calculate and answer 'no') or with the evaluative context (e.g., illegible text, laughter, emoticons, curse words). Both types of answer (i.e. noisy text and nonsensical responses) are defined as incoherent responses. A response will be said to be coherent if it is not incoherent. In the following, each type of incoherence will be defined in detail:

- *Question-independent incoherence*. Illegible answers are often detected as incoherent without having to know the question. This characteristic is one that is independent of the question. Likewise, the presence of faces, laughter or even bad-words are unacceptable in the evaluative context. These are therefore considered incoherent answers, regardless of the question. Some of these examples are detailed in Table 2. Another recurrent phenomenon in written responses is spelling and typing errors (e.g. phonemic errors, omitting letters, transposing consonants, pasting words, etc.). This feature is considered incoherent as long as the answers are illegible.
- *Question-dependent incoherence*. This type of incoherence is more sophisticated than the previous one since it requires information from the question in order to be detected. This type of incoherence typically occurs when the answer may be coherent for a different question than the one being asked, e.g. in table 3 the first answer "no" could be considered coherent for type 3 questions.

All 2019 responses were labeled as either coherent or incoherent in two stages. The first stage of labeling was performed by three teachers together with a member of the research team. Then, in the second stage, an analysis of the criteria used by each labeler was performed. Any labels that were misclassified according to said criteria were then corrected and the type of incoherence (i.e., dependent or independent of the question) was added. This second stage was necessary as each labeler in the first stage had a different perception (unsatisfactory agreement) of what they define as an incoherent answer. However, each of them managed to be consistent with their own definition. Thus, in the second stage, incoherence is robustly defined according to the criteria underlying the labels produced by each labeler. Additionally, we took care to ensure that there were no misclassified labels as this is an unbalanced data set.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 April 2022 doi:10.20944/preprints202204.0002.v1

Table 2. Examples of question-independent incoherent answers.

Answer	Observation
JDTGHKSLRJ	A characteristic type of incoherence is the random use of letters. In some cases, this is interpreted as laughter.
nooooo	Similar to laughter, some responses have words with elongated letters. This may be unintentional or to represent an exclamation.
7u7 :V XD	Emoticons are typographic representations of emotions and these are very popular in digital writing.
hi	There are key words in the answers, such as greetings and farewells.
a e i o u	Responses that do not contain any words are also recurrent.
+}{}{+´{-,.x	Another characteristic of incoherence is the excessive use of punctuation symbols.
bye jajajaja xd	Complex incoherent responses are those that use a mixture of keywords, laughter, faces and non-words.

[1] Originally in Spanish

Table 3. Example of each type of question with both coherent and question-dependent incoherent answers.

Type of question	Question	Answer	
		Coherent	Question-dependent incoherent
1	Maria and her husband cooked a tortilla yesterday, they divided it into 6 equal parts. Maria ate 2/6 and her husband ate 3/6. What fraction of the tortilla was left?	1/6	no
2	Catalina bought 12 onions. Of the 12 onions, she used 1/4 of them to make some delicious empanadas. How many onions did she use for the empanadas? Explain how you knew the result.	I need 3 and I know this because I divided 12:4=3x1=3	it is ok teacher
3	Camilo has to collect 60 balls. To find out how many balls he has left to collect, subtract 23 from 60. Is the exercise Camilo did correct? Justify your answer	it's ok because I added 37+23 and a half 60	43
4	Pablo takes 5 hours to travel from Santiago to La Serena. His friend Pedro traveled from La Serena to Santiago and took 300 minutes. Which of the two children took less time? Explain your answer	both took the same time because I multiplied 5x60=300 and 300 minutes is 5 hours	60x5 gives 300
5	What is a line of symmetry? Explain in your own words and give me an example	a line of symmetry is a line that separates two equal images	f
0	Pamela has 25 flowers and her friend gives her 17 flowers. Write in words the total number of flowers Pamela has	forty-two	areflowers

[1] Originally in Spanish

3.1.2. Data set description

This section is focused on the analysis of data for open-ended questions from 2017 and 2019. We have only labeled the data for 2019, with a total of 14,457 answers labeled according to types of coherence. Likewise, for the same year, 716 questions are labeled

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 April 2022 doi:10.20944/preprints202204.0002.v1

according to the six question types. It is important to clarify that for the data for 2017 is unlabeled. For this year, there are 1,180 questions and 16,618 answers, a higher number of questions and answers than for 2019. For more details see Table 4.

Table 4. Data for open-ended questions and answers from 2017 and 2019.

Year	Labeled	Questions	Answers
2019	True	716	14457
2017	False	1180	16618

Of the 716 questions in 2019, 30.16% of them are labeled as type 2 (Calculate with explaining). The number of type 1 questions are half as many type 2 questions. For more details see table 5. The next most popular question type is type 3 (Choose and/or affirmation), representing 29.18% of the questions. In smaller numbers, questions of type 0 (Others) are the least common.

Table 5. Data for tagged questions with the proportion by type of question from 2019.

Questions	Type of question (%)					
	0	1	2	3	4	5
716	3.77	15.92	30.16	29.18	9.77	11.17

Of the 14,457 responses from 2019, 13% are labeled as incoherent. Of these, 77% are question-dependent. For more details see Table 6. In terms of classification problems, incoherence detection is one of anomalous event detection, given its low presence compared to coherent responses. However, given the context of the platform, any percentage of incoherent responses is a worrying value.

Table 6. Data for tagged answers with the proportion by type of coherence from 2019. The notations C1 and C0 correspond to incoherent and coherent respectively.

Answers	C0 (%)	C1 (%)	Type of C1 (%)	
			Question-dependent	Question-independent
14457	86.66	13.33	77.17	22.82

3.1.3. Labeler agreement

Cohen's kappa is used to understand the degree of agreement between labelers (Table 7). The results do not reveal a significantly stable level of agreement when it comes to labeling responses based on coherence. Although a Cohen's kappa above 0.61 is considered substantial, our results do not return an overall value close to 0.91 among all labelers. The lowest level of agreement between two labelers was 0.68, while labeler 2 generally returned the lowest levels of agreement. On the other hand, the maximum values obtained are 0.91 and 0.95, both of which involved a member of the research team. However, when contrasting the labels with those corrected by us (note O in Table 7), Cohen's kappa reveals moderate to substantial agreement. Nevertheless, the values are still not above 0.81 (indicating a desirable degree of agreement). This may be because our labels are more robust for the definition of incoherence. By redefining incoherence taking into account the different criteria held by the labelers, then the degree of disagreement increased.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 April 2022 doi:10.20944/preprints202204.0002.v1

Table 7. Comparative table between all answer labelers. The notation O corresponds to our corrected labels, while the Lk notation corresponds to the kth labeler. That labeler A is above labeler B means that Cohen's kappa metric is determined between labelers A and B. Support corresponds to the number of responses labeling both labelers A and B.

Labeler	O				L1			L2		L3
	L1	L2	L3	L4	L2	L3	L4	L3	L4	L4
Cohen's kappa	0.68	0.56	0.70	0.67	0.73	0.81	0.95	0.68	0.72	0.91
Support	4130	4104	4149	14457	1964	1991	4130	4100	4104	4149

3.1.4. Proposed models

It is in our interest to fit a model that is able to determine whether a response associated with a certain question is incoherent or not. Such a model is a binary classification model where the positive class includes the incoherent answers. To solve this problem we consider two approaches: Single Model and Ensemble Model.

Both models follow the feature engineering paradigm, i.e., generate a manual representation of the data to be used in a prediction model (e.g., Perceptron). The reason for this assumption is that we need both models and interpretable features. Such a need comes from building predictors at the student level (e.g., the average proportion of numbers in a student's coherent answers).

The first approach is to use a single binary classifier with an attribute-based representation of the response and its associated question. This approach will be referred to as a Single Model.

Now, an important characteristic of the question-independent incoherent answers is that they only require the answer information to be detected. This quality can be used to better detect incoherence. If this type of incoherence were a known label then, subsequently, it would be sufficient for detecting question-dependent incoherence. In such a case, the type of question will be a information base before estimating for the detection of this more specific incoherence. In particular, answers to Calculate without explanation questions will be coherent if they at least have a numerical representation. Similarly, answers to Calculate with explanation questions will be coherent if they also have a sufficient number of appropriate words. This characteristic allows for the construction of incoherence classifiers specific to the type of question. However, classifiers that are capable of detecting question-independent incoherence and question type are needed. This approach will be referred to the Ensemble Model and each component will be detailed as follows:

1. *Question-independent incoherence detection*. With the question and answer the objective is to detect whether or not the incoherence is question-independent . This model also receives answers where the incoherence is question-dependent.
2. *Question type detection*. Given a question, the question type should be detected (e.g. Calculate without explaining).
3. *Question-dependent incoherence detection by question type*. Each type of question has a desired incoherent answer detector. With this, as many classifiers are required as the number of question types. In addition, these are models that only receive questions and answers where the incoherence is not question-independent.

In this way, the flow of the Ensemble model is as follows: (0) It receives an answer with its associated question; (1) the Question-independent incoherence detection model indicates whether or not the answer is of this type; (1.a) if it is, then the answer is incoherent; (1.b) otherwise, it goes to stage (2); (2) the Question type detection model indicates the type of question; (3) with this information, the Question-dependent incoherence detection model is used to detect whether or not the answer is incoherent.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 April 2022 doi:10.20944/preprints202204.0002.v1

3.1.5. Question type detection

Question type detection is a particular form of text classification within natural language processing (NLP) [15] . In our case, we consider six types of question in order to detect what type a new question is, using only information from the question itself. For text classification, various techniques have been used, ranging from Machine learning with Bags-of-words [16] to Deep learning [17].

For classifying question types we will use a sentence representation studied by [1] based on grammatical attributes and similar to the work of [18] based on syntactic and lexical attributes. For this, we will consider POS and dependency tags for the question using the Spanish version of the Spacy library (Authors on https://explosion.ai/, accessed on March 16, 2022).

A second proposed model is based on a representation of the questions using the BERT model of language in Spanish (BETO). This is a representation that has obtained outstanding results in text classification when compared to traditional Machine learning [19]. There are even other works using the same technique [20], [2].

For the predictive model, the Support-vector classifier will be applied. This is a classifier used by [3] in the prediction of question types.

3.1.6. Engineering and selection of features

- *Traditional attributes*. These are Single attributes that only use the information of the tokens in the answers. These are divided into two groups, some at answer-level and others at token-level. These attributes are detailed below.
  The following attributes are considered at the answer level:
  - *Length*. This consists of the number of characters in the answer (not including spaces). For answers to questions that ask for an explanation, the length of the answer is important. In this sense, very short answers may be considered incoherent. A similar characteristic is the number of tokens in the answer, though answers with many words may be of the same length as answers with few words.
  - *Punctuation marks*. These are typographic symbols other than numbers and letters, which help the written text (e.g. ,.;+|%#). In mathematics, they play a fundamental role in numerical representation (e.g. fractions and decimals) and operations (e.g. * for multiplication and / for division). In other cases, rare symbols such as ! {) may indicate incoherence.
  - *Alphabetics*. These are the letters in the answer. A useful characteristic is the number of letters that are vowels. If the proportion of vowels is extreme (either very small or very large) this may be an indicator that the response includes non-words (e.g. ghjghj representing laughter).
  - *Numbers*. Most answers have numbers to represent quantities, equations and explain operations. In some cases, numbers that are too large (e.g., 123456789) may indicate incoherence. Likewise, the presence of appropriate numbers in questions that ask for calculations may indicate that the answer is coherent. However, some students use words for the numbers (e.g., thirty-seven) and sometimes they make spelling mistakes. In this case, the triangular matrix algorithm [13] is used to correct and transform numerical representations (e.g. five ogtaves to 5/8).

  In addition, token-level attributes are also considered, which allow for finer characterization of non-words. These include:

  - *Character repetition*. This consists of detecting the consecutive repetition of letters in a token. In Spanish it is uncommon to repeat the same letter more than twice, unless they are exclamations (e.g. noooo where letter o is repeated four times).
  - *Character frequency*. This time the frequency of letters in a token is calculated, though not necessarily consecutive. It is rare to find words with certain consonants (e.g. k, w, ñ) more than once in a Spanish word.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 April 2022 doi:10.20944/preprints202204.0002.v1

- *Semantic attributes*. These consist of attributes that consider the meaning of some tokens and phrases. Some tokens represent faces (e.g. xd) and others do not belong to official dictionaries (e.g. lol). We detail each of these, below:
  - *Dictionaries*. Two dictionaries are used to highlight some of the words in the answer. For words that are spelled correctly, the Real Academia Española (RAE) is used (Authors on https://pypi.org/project/pyrae/, accessed on March 16, 2022). While for some colloquial tokens the UrbanDictionary is used (e.g. uwu, lol, xd) (Authors on http://api.urbandictionary.com/v0, accessed on March 16, 2022). The proportion of tokens in each dictionary can indicate the degree of coherence.
  - *Faces*. The presence of emoticons (e.g. :), *_*) is more common in responses that are indicated as incoherent. Nowadays, these have been augmented by emojis [50], graphical representations of emotions (e.g. :smile-face: to smile face).
  - *Keywords*. There are some words that are not available in dictionaries and others that have a negative value. The former corresponds to slang (e.g. dunno, c'mon), while the latter corresponds to curse words.
- *Contextual attributes*. These attributes are based on the intersection between answer and question tokens. For certain question types, tokens from the question are also needed in the answer, e.g. character names. Attributes such as this and others are detailed, below:
  - *Binary words*. Some question types require certain key words in the answers. The most common are yes or no. Some of these words are replaced by right, good, correct or affirmative, and their negations. The presence of any of these is detected in the answer.
  - *Key questions*. There are questions that require a finer distinction than just the six main categories. For this, we check for the presence of certain keywords in the question. These include: To be & (right | wrong), To be & (possible | impossible). For morpheme variations (e.g. time, quantity, gender) their variants are joined.
  - *Overlap*. A Single way to compare similarity between question and answers is to count the common tokens. This can be very coarse, as some variations in the tokens may differ between two tokens but refer to the same thing (e.g. plurals). For this, soft intersection is considered, i.e., pairs of tokens in question and answer with sufficient similarity using Levenshtein distance [12] (in our experiments greater than 0.8). For certain types of questions the presence of special question words in the answer is relevant. The detection of proper noun (PROPN) and nominal subject (nsubj) in the question is considered. For POS tags (e.g. PROPN) and dependency tags (e.g. nsubj) the Spanish version of Spacy library is used (Authors on https://explosion.ai/, accessed on March 16, 2022).

In addition to the manually designed attributes, the embedding components of both the answer and its associated question are added using the BETO language model.

Each of these attributes must be selected for a given classifier model. In this case, we have seven binary classifiers, one for question-independent incoherence classifier and one question-dependent incoherence classifier for each question type.

The following scheme is used for feature selection:

1. *Feature ranking*. chi2 is used to assign a relevancy value to each feature. This technique has been studied in other text classification problems with effective results [14].
2. *k-best features*. For a fixed k the first k-best features are chosen. Then, an optimal k is such that it maximizes the performance for the test set.
3. *Validation*. In order to avoid overfitting we use fifty 5-fold cross validation.

3.1.7. Evaluation

The following evaluation scheme is proposed for evaluating each detection model:

 doi:10.20944/preprints202204.0002.v1

1. *Metrics*. The problem of classifying responses according to coherence is made difficult due to a significant imbalance in the number incoherent responses (they only account for 13% of the responses). This imbalance causes metrics such as accuracy to be inappropriate. For example, if the proportion of well-classified versus total data (accuracy) is considered as a metric, then the classifier that always predicts the majority class returns results that are close to the imbalance ratio (in our case 90.9%). Therefore, two descriptive metrics of false positives and false negative are used, which are Precision and Recall, respectively. Instead of accuracy, the harmonic mean between Precision and Recall (F1 score) is used. In addition, the data proportions for each class (Support) will be indicated.
2. *Validation*. To validate, the k-fold cross validation technique is used. This consists of randomly dividing the data set into k chunks of similar sizes. Of the k chunks, one is used for validation as a test set and the remaining k-1 chunks are used for training. The process is repeated k times, with one chunk for used validation each time. Training and testing with the entire data set allows us to have an estimate with k samples of the performance of a classifier. This is more representative than an estimate using only one sample and following the classical approach. For the same reason, N random repetitions of k-fold cross validation are performed, this corresponds to Nk fittings. Additionally, the data set is not randomly split, since it could be the case that two answers associated to the same question are left in training and testing. Given this, a stratified randomization of the data is performed based on the question. As a result of this, no answers to the same question are included in different sets, thus avoiding data contamination (in our experiments N=50, k=5).
3. *Baselines*. Two basic approaches will be used to compare these models. The first one consists of the predictive ability of human labelers, i.e. the degree of agreement between labelers for the detection of incoherence (see subsection X.). We find that the labelers have an average Cohen's kappa of 0.80, a minimum value of 0.68 and a maximum value of 0.95. In order to be able to compare we will use the metrics for the detection of incoherence. For the sake of comparison we will also use the metrics already proposed, where an average F1-score of 0.82, a minimum value of 0.72 and a maximum value 0.96 are obtained. The second is a binary classifier and based on a language model to represent the sentences. For the language model we use BETO [6], inspired by BERT [7] but pre-trained with Spanish text. And for the binary classifier, Support-vector classifier (SVC) with RBF kernel is selected.

3.2. Predicting the scores on an end-of-year national standardized test

3.2.1. Data

We focused on using data gathered from the ConectaIdeas platform during 2017 from both open-ended and closed-ended exercises. Fourth grade students used this platform to exercise mathematics content. Additionally, in order to measure progress in terms of performance when using the platform, 24 schools participated in a pilot program. In this pilot, only one of the two classes in each school worked on the platform (treatment group), while the other class did not have any access to the platform (control group). We will only focus on the responses given by students in the treatment group, since they are the only ones who answered both open- and closed-ended questions during the year.

Now, for simplicity, we only consider two important annual tests, one at the beginning of the year and the other at the end of the year. At the beginning of the year, a test was applied to measure the performance of students prior to participating in the program. This test is called the pre-test. The second test, called SIMCE, is a national standardized test that is conducted every year by an educational quality agency. We consider the pre-test per student as an estimator to predict the score on the end-of-year national standardized test (SIMCE). It is worth noting that, at a national level, this standardized test has a mean of 260.95 and a standard deviation of 47.8, both of which are important reference values [36].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 April 2022 doi:10.20944/preprints202204.0002.v1

Additionally, we include the average SIMCE score obtained by the school in 2016 as another estimator. Note that this estimator is at the school level and not per student. Also, we automatically label responses to open-ended questions with two prediction models, tuned as detailed in this paper. The first is a question type detector, which consists of receiving a question and responding with one of six question types. The second is a coherent answer detector, which receives both the answer and the question in order to detect whether or not the answer is incoherent.

3.2.2. Data set description

Below, we will perform a general analysis of the data for the detection of 2017 SIMCE scores. In our case, we are interested in all those students on the platform who have answered at least one open-ended question (these are the written answers). In summary, Table 8 reports that only 58.29% of the students out of a universe of 796 students from across 24 schools have answered this type of question at least once. This corresponds to 464 students. Additionally, we do not have a balance between male and female students, with only 44.59% corresponding to female students. It can also be observed that neither the sex of the student nor the number of students we are interested in are stable at the school level, since there are schools with more girls than boys (school 7 in Table 8) and others with 78.04% (school 13 in Table 8) of students with responses to open-ended questions.

Table 8. Summary of students considered by school. These are separated by sex. Additionally, The Support column corresponds to the percentage of students who answered at least one open-ended question.

School	Students			
	Total	**Male** (%)	**Female** (%)	**Suppport** (%)
1	38	63.15	36.84	68.42
2	37	54.05	45.94	64.86
3	20	65.00	35.00	65.00
4	21	71.42	28.57	38.09
5	39	56.41	43.58	66.66
6	26	50.00	50.00	65.38
7	30	40.00	60.00	83.33
8	31	54.83	45.16	48.38
9	35	68.57	31.42	48.57
10	44	59.09	40.90	50.00
11	32	56.25	43.75	62.50
12	24	45.83	54.16	37.50
13	41	60.97	39.02	78.04
14	35	57.14	42.85	71.42
15	48	50.00	50.00	52.08
16	32	65.62	34.37	65.62
17	36	52.77	47.22	50.00
18	31	45.16	54.83	45.16
19	29	48.27	51.72	68.96
20	27	55.55	44.44	59.25
21	39	53.84	46.15	48.71
22	28	53.57	46.42	35.71
23	35	48.57	51.42	40.00
24	38	55.26	44.73	73.68
Summary				
24	796	55.40	44.59	58.29

Now, if we filter by all students with at least one response to an open-ended question, Table 9 summarizes the number of students per school. When we reduce the sample to this universe of students, we obtain a more even distribution between male and female students (51.5% and 48.49%, respectively). It should be added that the 24 schools we studied

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 April 2022 doi:10.20944/preprints202204.0002.v1

correspond in general to at-risk schools, with a high score on the school vulnerability index. The weighted average of this index (according to the number of students per school) for our data is 0.902 of vulnerability. On the other hand, in comparison to the national SIMCE of the same year, most of the schools are below the national average with almost 0.381 deviations from SIMCE standards. Despite this, there are students in some schools that stand out with 2.45 SIMCE standard deviations above the national average.

Table 9. Summary of students considered by school. These are separated by sex. Additionally, Support column corresponds to the percentage of students who answered at least one open-ended question. The average (avg), minimum and maximum per school are added. Additionally, the IVE column corresponds to the School Vulnerability Index (IVE, for its acronym in Spanish).

School	**Students**			**SIMCE** (std SIMCE)			**IVE**
	Total	**Male** (%)	**Female** (%)	**avg**	**min**	**max**	
1	26	69.23	30.76	0.187	-1.533	1.723	0.943
2	24	54.16	45.83	-0.733	-2.462	1.041	0.944
3	13	53.84	46.15	-1.424	-2.635	0.553	0.944
4	8	75.00	25.00	-0.669	-1.922	1.218	0.913
5	26	53.84	46.15	0.086	-1.312	2.452	0.918
6	17	41.17	58.82	-0.470	-1.552	0.928	0.937
7	25	40.00	60.00	0.017	-2.578	1.878	0.906
8	15	60.00	40.00	-0.844	-2.123	0.755	0.763
9	17	58.82	41.17	-0.496	-2.688	1.599	0.752
10	22	59.09	40.90	-0.744	-2.594	0.749	0.797
11	20	50.00	50.00	-0.638	-2.283	0.620	0.769
12	9	77.77	22.22	-0.946	-2.113	0.135	0.823
13	32	53.12	46.87	0.961	-0.949	2.355	0.871
14	25	48.00	52.00	-0.655	-2.374	1.667	0.943
15	25	36.00	64.00	-0.396	-1.874	1.510	0.967
16	21	52.38	47.61	-0.143	-1.686	1.986	0.923
17	18	50.00	50.00	-1.075	-2.475	0.464	0.913
18	14	42.85	57.14	-0.605	-2.582	1.041	0.961
19	20	40.00	60.00	-0.313	-2.492	1.395	0.910
20	16	50.00	50.00	-0.991	-2.129	1.542	0.937
21	19	47.36	52.63	-1.056	-2.989	0.820	0.948
22	10	60.00	40.00	-0.478	-1.868	1.074	0.955
23	14	42.85	57.14	0.211	-0.630	1.718	0.947
24	28	50.00	50.00	-0.213	-1.925	1.471	0.940
Summary							
24	464	51.50	48.49	-0.381	-2.989	2.452	0.902

Finally, to recall, in 2017 we have 1,180 open-ended questions and 16,618 answers to this type of question. When compared with the number of answers to closed questions, we see that almost 36.4 times more closed questions are answered than open-ended questions. Similarly, almost 2.8 times more closed questions are asked than open-ended questions. For more details see Table 10. Students therefore answer, on average, 36.4 times more closed-ended questions than open-ended questions. For more details see Table 11.

Table 10. Summary of number of answers to Open-ended and Close-ended questions.

Questions		**Answers**	
Open-ended	Close-ended	Open-ended	Close-ended
1180	3161	16618	621575

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 April 2022 doi:10.20944/preprints202204.0002.v1

Table 11. Average of answers per student.

Open-ended	Close-ended
35.81	1303.78

3.2.3. Open-ended model

We collected a series of answers per student to both open-ended and closed-ended questions. Answers to open-ended questions are text written by students in a digital platform, while answers to closed-ended questions are not. We will focus on building a model that uses the written responses to predict a final score. The way students write their responses to mathematical problems has been studied and investigated as predictors of writing quality [45]. Other authors have studied the connection between writing style and students' understanding of mathematical problems [44]. These studies have examined the previously unexplored territory between text and mathematical thinking [43]. A fundamental property of these responses is that they are digitally written texts, a feature that brings its own challenges, e.g. the detection of coherent answers given a question type. We use question type detection and coherence detection models to automatically tag our data. With this, we can filter the answers according to answer types.

On the other hand, several models have been tested for predicting a final score [40], [39], [48]. We consider a linear model for predicting scores, such as the models used by [36], [35] and [49]. For this a student i is represented with a vector of regressors $x_i = (x_{i,0}, x_{i,1}, ..., x_{i,k})$ and its score is defined by:

$$\widehat{\text{score}}_i(b, w) = b + w_0 \cdot x_{i,0} + w_1 \cdot x_{i,1} + ... + w_k \cdot x_{i,k}$$

where the slope $w = (w_0, w_1, ..., w_k)$ and the intercept b are parameters of the fitted least squares regression with n students, that is:

$$\begin{cases} \text{minimize} & \frac{1}{2n} \cdot \sum_{i=1}^{n} (\text{score}_i - \widehat{\text{score}}_i(b, w))^2 \\ \text{s.t.} & w \in \mathbb{R}^k, b \in \mathbb{R} \end{cases}$$

Now, traditional models use demographic and process variables as regressors for the student. This time, we are interested in using written responses to open-ended questions as regressors of the linear model. To do this, there are several challenges in working with unstructured data such as text. In particular, how to condense the information from various answers to different types of questions in different periods of the year into a one-dimensional representation. The simplest way is to consider only one response and retrieve attributes from it in order to construct variables in the linear model. Studies such as [41] design a set of regressors from the text in order to predict a continuous variable. Others use teacher comments to predict student scores [4], [5]. Additionally, we deal with a series of responses and transform them into a useful predictor. For this, the following section is devoted to a regressor design based on written responses to open-ended questions. The model that uses this type of regressor will be referred to as an Open-ended model.

3.2.4. Engineering and selection of regressors

- *Traditional*. According to the study by [36] we will consider two groups of regressors:
  - *Historical*. These correspond to variables that are expected to remain unchanged during the year. Among these are the grade, school, sex, SIMCE score of their school in the previous year, and school vulnerability index (IVE).
  - *Dynamic*. Unlike the previous ones, these variable can vary during the year, e.g. number of exercises answered on the first attempt. In this context these variables are usually called Process variables.
- *Based on written answers*. One of the big challenges is to capture relevant information in the written responses to build predictors of a final score. The simplest way to do

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 April 2022 doi:10.20944/preprints202204.0002.v1

this is to count how many such responses students make per year. However, this way of doing it omits valuable information in the words they use and the structure of students' answers. Of course, not all answers are appropriate for this, e.g. in some cases incoherent answers may discarded. Also, answers to certain types of questions have tokens of interest, e.g. numbers accompanied by units of measurement. We describe how we tackle this issue, below:

- *Simple indicators*. These are indicators that allow data to be aggregated. In our case, to aggregate the answers to open-ended questions. One indicator is the number, e.g. number of incoherent answers. Another one is the proportions, e.g. the proportion of coherent answers. Also, double aggregate indicators, e.g. number of coherent answers to Calculate with explaining questions.
- *Traditional/Semantic/Contextual features*. Different features can be extracted from any given answer. In the section on detecting the incoherence of an answer, we detailed some of the features that were separated into three groups. These are Traditional, Semantic and Contextual features. Given all the answers that are given by a student, the way in which these features are distributed is important, e.g. Average of number of words in coherent answers. These features can also be aggregated by question type, e.g. Standard deviation of number of numbers in answers to Calculate without explaining questions.
- *Linguistic features*. Unlike the previous features, this time we are interested in using linguistic knowledge to capture useful information from the answers. The most abstract features of an answer are (shape) the shape of the tokens and (alpha) indicator of whether the token is alphabetic. An example of this is the answer [There, are, 25, candies] with (shape) [Xxxxx, xxx, dd, xxxxxxx] and (alpha) [1, 1, 0, 1]. Some words are noted for their abundance, these are called stop-words, e.g. [There, are] are of this type but [25, candies] are not. Shape, alpha and stop-word features are easy to detect, since in particular they do not require information from the other tokens. However, they can condense valuable information about a student's answers, e.g. Average of number of stop-words in answers to Calculate with explaining questions.
  We will now consider two more attributes, but this time they depend on the other words and need to be detected automatically. These are part-of-speech tagging (POS tags) and dependencies (dep), both of which are available in the Spanish version of the Spacy library. POS tags are detected for each token in a response and can be: (PRON) Pronoun, (ADJ) Adjective, (VERB) Verb, (NUM) Numeral, among others (Available in https://universaldependencies.org/u/pos/, accessed on March 16, 2022). An example of a regressor using POS tags is the following: Median number of verbs (tag/VERB) in coherent answers. Dependency tags correspond to syntactic dependencies between the tokens of the answer, e.g. If we remove [candies] from the response [There, are, 25, candies] then 25 will take the role of the root (dep/ROOT) of the sentence that had [candies]; Likewise, [25] loses its characteristic of being a number with units, a dependency called Numeric modifier (dep/nummod). An example of a regressor using dependency tags is the following: Average of number of dep/nummod in coherent answers. Other interesting dependency tags include: (obj) object, (nsubj) nominal subject, (nmod) nominal modifier, among others (Available in https://universaldependencies.org/u/dep/, accessed on March 16, 2022).

All of these feature-based regressors can be applied to a subset of responses and condensed into one dimension using specific functions. Further details of this are provided, below:

- *Filters*. The simplest way to construct a regressor based on the answers is to consider all of them. Another way is to use a subset of the answers, e.g. only incoherent answers. This way of filtering is based on coherence. Also, the type

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 April 2022 doi:10.20944/preprints202204.0002.v1

of question associated with the answer is automatically detected, so that it can be grouped according to question type, e.g. only answers to questions of type Others. Likewise, both filters can be applied, i.e. both according to coherence and question type, e.g. only incoherent answers to questions of type Others.

- *Functions*. We only focus on six functions on the features, these are: sum, average, standard deviation, minimum, maximum and median. For example, if we select the attribute number-of-tokens in an answer we can construct regressors by summing across all of the answers the number of tokens in the answer, if we want to see the expected value of the number of tokens just average the number of tokens, if the dispersion of this attribute is of interest we use the standard deviation. For other attributes it is useful to calculate the maximum, minimum and median across all answers. Likewise, it is also possible to look at a subset of answers, e.g. Minimum number-of-tokens in incoherent answers to Calculate without explaining questions.

The following scheme is used for selecting regressors:

1. *Filtering*. All regressors with absolute correlation less than 0.19 are discarded. This is to avoid choosing regressors that are noisy but useful.
2. *Genetic algorithm*. An algorithm that tests all combinations of regressors is impractical since the number of candidates is exponential in the number of regressors. For this reason, and simplicity's sake, we rely on a genetic algorithm for selecting regressors. Based on the work of [37], the algorithm consists of the following stages: (0) Initial population; (1) Population evaluations; (2) Reproduction: select k individuals from the current population; (3) Mutation of the k individuals; (4) Cross-over between individuals and the current population; (5) Evaluation and selection of the fittest individual; and (6) Return to (1) and repeat the process. In our experiments: (0) Baseline regressors; (1) R2 as evaluation metric; (2) Three individuals; (i1) randomly selected, (i2) best R2 and (i3) worst R2; (3) Mutating corresponds to removing or adding regressors until R2 does not improve; (4) Cross-over corresponds to keeping common regressors and randomly choosing regressors that are not common; (5) Select individual with higher expected R2; (6) Repeat but stop if R2 does not increase.
3. *Reduction*. Two things happen: (a) Too many regressors tend to be selected in relation to the number of regressors in the baseline model, and (b) linear models with more regressors tend to predict the observed variable better. To address this, we reduce the number of regressors selected until we obtain a similar number of regressors as the baseline model. The reduction is performed as follows: (1) From the selected regressors one is chosen and R^2 is calculated without this chosen regressor; (2) The regressor that improves the R^2, or decreases it as little as possible, is eliminated; (3) The selected regressors are updated and the process is repeated from (1) until as many regressors as the baseline model are obtained.
4. *Validation*. In order to avoid overfitting we use two hundred and fifty 4-fold cross validation.

3.2.5. Evaluation

The following evaluation scheme is proposed to evaluate the model:

1. *Metrics*. For regression models we use two statistical estimators to measure the model fitting. The first is the coefficient of determination or R2, an indicator between 0 and 1 where closer to 1 is better. The second is the root mean square error or RMSE, an indicator that is on the scale of the observed variable where smaller is better. We will usually consider a normalized version of the RMSE in terms of the standard deviation of the observed variable at the national level (std SIMCE), simply divide RMSE with std SIMCE. The standard deviation corresponds to 47.80 according to the work of [36].
2. *Validation*. To validate, the k-fold cross validation technique is used. This consists of randomly dividing the data set into k chunks of similar sizes. Training and testing with the entire data set allows to have an estimate with k samples of the performance

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 April 2022 doi:10.20944/preprints202204.0002.v1

of a classifier more representative than an estimate with only one sample when using the classical approach. For the same reason, N random repetitions of k-fold cross validation are performed, this corresponds to Nk fittings. Additionally, the data set is not randomly split so as to avoid two students from the same establishment being left in training and testing. To do so, a stratified randomization of the data is performed based on establishment, ensuring that no students to the same question are in different sets, thus avoiding data contamination (in our experiments N=250, k=4).

3. *Baseline*. To compare the model based on open-ended questions we propose a single baseline. As with the proposed model, the baseline is a linear regression model. This time, the regressors are the so-called traditional regressors, separated into historical and dynamic regressors. This model consists of 14 regressors: a double regressor corresponding to whether the student is Male or Female, a single socio-economic regressor named School vulnerability index (IVE), and other regressors come from exercises completed during the year (or Process variables).

4. Results

4.1. Detecting coherence of answers to open-ended questions

For question type detection, we tested two prediction models. The first one uses linguistic attributes (denoted by L) and the second one uses question embedding with BETO (denoted by B). Overall, model B is superior to model L in F1-score for the detection of all question types (types 0 to 5). See Table 12 for details. Both multiclass models have a satisfactory performance for the detection of questions of type 2, 3, 4, 5, but not for questions of type 0 and 1. In particular, the fact that model B outperforms model L for questions of type 0 (Others) may be due to the fact that the diversity of this type of question is generalized through word embedding using the language model BETO, but not the linguistic features.

Table 12. Summary of Precision, Recall, F1-score and Support metrics for each Question type detection model in test set. These are obtained from fifty 5-fold cross validations. The notation L and B denotes each multi-class model these are Linguistic features and BETO features respectively. Each row corresponds to the metrics for these classes with the multi-class models.

Class	Precision		Recall		F1-score		Support
	L	B	L	B	L	B	
0	0.96	**0.99**	0.36	**0.49**	0.51	**0.64**	9.0
1	0.86	0.86	0.85	**0.95**	0.85	**0.90**	38.2
2	0.88	**0.93**	0.94	**0.99**	0.91	**0.96**	71.6
3	0.91	**0.97**	**0.98**	0.97	0.94	**0.97**	69.2
4	0.98	**0.99**	0.89	**0.90**	0.94	0.94	22.8
5	0.97	**0.98**	0.94	**0.95**	0.95	**0.97**	26.0

We will now study the results obtained from the feature selection to seven binary classifiers using only the attributes obtained with BETO. These classifiers are: (I) Question-independent incoherent, (DQ) Question-dependent incoherent by type of question Q. First, unlike the previous problem, this time an attribute selection with chi2 is performed for all binary classification problems. Table 13 summarizes the number of attributes selected and the proportion of these that use only the answer (denoted by Answer) or only the question (denoted by Question). In order to detect the independent incoherence of the questions, 371 attributes are needed. 99.2% of these correspond to attributes of the answer, which is consistent with the type of incoherence. In addition, incoherence detection for other question types such as 0 and 5 have a balanced distribution between question and answer attributes. This is not the case for question types 1 and 3, where there is little presence of useful information for detecting question-dependent incoherent answers. See Table 13 13 for details.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 April 2022 doi:10.20944/preprints202204.0002.v1

Table 13. Description by type of features for the models based only on embeddings with BETO of the question and answer. If the selected attributes use the answer then they are denoted by Answer and if they use the question then they are denoted by Question. Each row corresponds to a binary classification model. These classifiers are: (I) Question-independent incoherent, (DQ) Question-dependent incoherent by type of question Q. BETO model corresponds to the baseline that uses only features with BETO.

Classifier (B)	**Number of features**	**Answer** (%)	**Question** (%)
I	371	99.2	0.8
D0	118	45.8	54.2
D1	179	84.9	15.1
D2	700	73.7	26.3
D3	405	92.8	7.2
D4	614	41.4	58.6
D5	738	67.1	32.9
BETO model	491	91.0	9.0

On the other hand, we will study the results obtained from the feature selection with the same seven binary classifiers, but this time using the same three main types of attribute (Traditional/Semantic/Context) as those obtained with BETO. This time most of the classifiers select fewer attributes than the BETO-only models (see Table 13). This may be due to the fact that the manually designed attributes may capture the important information in less dimensions than when using BETO. A different case is observed with the classifier for question type 0 (Other), which increases the number of selected attributes. Another phenomenon observed from Table 14 is that most of the selected attributes correspond to those that only use information from the answer. Likewise, compared to the models that only use BETO, the proportion of attributes that use the question in some cases is reduced to less than half. See Table 14 for details.

Table 14. Description by type of features for the models based on mixed features of the question and answer. If the selected attributes use the answer then they are denoted by Answer and if they use the question then they are denoted by Question. Each row corresponds to a binary classification model. These classifiers are: (I) Question-independent incoherent, (DQ) Question-dependent incoherent by type of question Q. Single model corresponds to the baseline that uses mixed features.

Classifier (M)	**Number of features**	**Answer** (%)	**Question** (%)
I	94	97.9	2.1
D0	337	45.4	54.6
D1	78	85.9	14.1
D2	73	97.3	2.7
D3	81	96.3	3.7
D4	11	72.7	27.3
D5	219	97.3	2.7
Single model	671	80.6	19.4

We will now describe the analysis performed using the mixed attribute models. This time we analyze the proportion of attributes according to the type of attribute: Traditional, Semantic, Context or BETO (see Table 15 for details). In general, all classifiers select at least 50% of the component attributes of the question or answer vectors with BETO, except the classifier for question type 4 (Compare quantities) with zero attributes coming from BETO. The same classifier has 27% of the Context attributes, and 63.6% of the Traditional attributes. These percentages may be due to the small number of attributes selected (only 11). Additionally, in general, the Traditional attributes are the most predominant after those with BETO in most classifiers. This may be due to the large number of attributes of this type that are selected. Classifiers for detecting incoherence in question types 3 (Choose and/or affirmation) and 4 (Compare quantities) are the only ones that have a higher percentage of Context attributes. This is consistent with the definition of this type of question, since

 doi:10.20944/preprints202204.0002.v1

answers of this type will be coherent if they share specific tokens with the question (e.g. personal pronouns). Classifiers for detecting incoherence in question types 0 (Others) and 5 (Procedure and content knowledge) are the only ones that have more than 70% of BETO attributes. This may be because the former are too diverse and the latter require content knowledge in order to detect coherent answers.

Table 15. Description by type of features for the models based on mixed features according to the three types of features (Traditional/Semantic/Context) plus embedings with BETO. If the selected attributes use the answer then they are denoted by Answer and if they use the question then they are denoted by Question. Each row corresponds to a binary classification model. These classifiers are: (I) Question-independent incoherent, (DQ) Question-dependent incoherent by type of question Q. Single model corresponds to the baseline that uses mixed features.

Classifier (M)	**Number of features**	**Traditional** (%)	**Semantic** (%)	**Context** (%)	**BETO** (%)
I	94	27.7	14.9	4.3	53.2
D0	337	7.7	5.3	1.2	85.8
D1	78	29.5	14.1	5.1	51.3
D2	73	28.8	16.4	4.1	50.7
D3	81	30.9	11.1	6.2	51.9
D4	11	63.6	9.1	27.3	0.0
D5	219	13.7	6.8	1.4	78.1
Single model	671	6.1	3.0	1.3	89.6

To conclude, we want to know which of the classifiers described above has a better predictive ability. For this, the models with mixed attributes are denoted with the letter M, while the models that only use BETO attributes are denoted with the letter B. All the results are quantified in Table 16. In summary, binary classification models that use both BETO-captured information and manually designed attributes make better predictions than those that only use BETO attributes. These results are significant, except for the detection of dependent inconsistency in questions of type 0 (Other) and type 5 (Procedure and content knowledge), where the differences in F1-score are not as drastically outstanding as those of the other binary classifiers. This may be due to the fact that detecting independent incoherence and dependent incoherence in questions of type 1, 2, 3 and 4 is easier than with type 0 and 5. This is due to the simplicity of the incoherence criteria.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 April 2022 doi:10.20944/preprints202204.0002.v1

Table 16. Summary of Precision, Recall, F1-score and Support metrics for each model in test set. These are obtained from fifty 5-fold cross validations. The notation M and B corresponds to Mixed features model and BETO features model respectively. The notations C1 and C0 correspond to incoherent and coherent respectively. The prefixes i and d in C1 corresponds to Question-dependent incoherent and Question-independence incoherent respectively. Each row corresponds to a binary classification model. These classifiers are: (I) Question-independent incoherent, (DQ) Question-dependent incoherent by type of question Q.

Classifier	Class	Precision		Recall		F1-score		Support
		M	B	M	B	M	B	
I	C0 | C1-d	0.90	0.90	0.90	0.90	0.90	0.90	2803.4
	C1-i	0.91	0.91	**0.90**	0.71	**0.90**	0.79	88.0
D0	C0	**0.93**	0.90	0.96	0.96	0.94	0.94	79.6
	C1-d	**0.66**	0.63	**0.63**	0.60	**0.59**	0.57	11.0
D1	C0	**0.99**	0.97	0.99	0.99	**0.99**	0.98	369.2
	C1-d	**0.96**	0.87	**0.93**	0.47	**0.94**	0.60	18.6
D2	C0	**0.98**	0.96	**0.99**	0.96	**0.98**	0.96	813.4
	C1-d	**0.94**	0.83	**0.92**	0.83	**0.93**	0.83	149.0
D3	C0	**0.98**	0.96	**0.96**	0.95	**0.97**	0.96	681.0
	C1-d	**0.81**	0.78	**0.92**	0.79	**0.86**	0.79	131.2
D4	C0	**0.99**	0.97	**0.95**	0.93	**0.97**	0.95	278.4
	C1-d	**0.83**	0.70	**0.98**	0.82	**0.90**	0.75	49.2
D5	C0	0.97	0.97	0.98	0.98	**0.98**	0.97	284.2
	C1-d	**0.81**	**0.80**	0.78	0.75	**0.79**	0.77	26.6

Given this, all of the individual classifiers (including both independent incoherence and dependent incoherence by question type) are replaced with their mixed-attribute version. This is because the mixed-attribute classifiers make better predictions than the BETO-only attributes. The Ensemble model is therefore fixed and consists of the mixed-attribute model. We now have to compare the incoherence models. These are: (E) Ensemble Model, (S) Simple Model and (B) BETO Model. Before this, Table 13 shows that model B has 491 attributes and uses 91.0% of the answer attributes. On the other hand, Tables 14 and 15 show that model S has 671 attributes, of which 89.6% correspond to BETO attributes and 80.6% to answer attributes.

Finally, models E, S and B were subjected to the same evaluation criterion: fifty 5-fold cross validations stratified by question. Table 17 summarizes the results of the main task of detecting the coherence of answers to open-ended questions. Model E successfully solves the task with a better performance than human labelers (model H), with an F1-score of 0.86 versus 0.82 for model H. Additionally, the baseline model with mixed attributes manages to outperform the baseline model with only BETO attributes. However, both are one tenth of an F1-score below the model E.

Table 17. Summary of Precision, Recall, F1-score and Support metrics for each Coherence detection model in test set. These are obtained from fifty 5-fold cross validations. The notation E, S and B corresponds to the Ensemble model, Single model and BETO features model, the Human performance is denoted by H. The notations C1 and C0 correspond to incoherent and coherent respectively.

Class	Precision				Recall				F1-score				Support	
	E	S	B	H	E	S	B	H	E	S	B	H	*	H
C0	0.98	0.96	0.96	0.98	0.97	0.96	0.95	**0.98**	0.97	0.96	0.95	**0.98**	2510.9	3073.5
C1	**0.83**	0.76	0.71	0.80	**0.90**	0.79	0.75	0.80	**0.86**	0.78	0.73	0.82	380.4	332.7

[1] * The support for the classifiers is the same.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 April 2022 doi:10.20944/preprints202204.0002.v1

4.2. Predicting the score on an end-of-year national standardized test

Before showing the Open-ended model with its selected regressors, we will first show other regressors that use the answers to open-ended questions and that were not selected, despite having a significant correlation. First, Table 18 reports a series of regressors, based on Simple indicators and Traditional/Semantic/Context features of the answer. Each of these are determined using different filters by answer type and different types of functions (see Filters and functions in regressor engineering section). One of the most important regressors is probably the number of responses, with a correlation of 0.34 with the SIMCE score. However, if only coherent responses are considered then the correlation improves to 0.43. A similar attribute is the proportion of coherent responses per student, with a correlation of 0.42. However, if we only consider the answers to type 2 questions (Calculate with explaining) then the average length of an answer has a correlation of 0.45 with the SIMCE score. This positive ratio indicates that students with a higher average answer length to type 2 questions obtain better scores, while those with a lower average answer length obtain worse scores. Another regressor with a strong correlation is the total amount of tokens that the student uses in their answers, with a correlation of 0.48. In this sense, the more tokens they use the higher their SIMCE score. For other regressors, see Table 18.

Table 18. Some interesting regressors using Simple indicators and Traditional/Semantic/Context features of the answer. Correlation, corresponds to the Pearson coefficient between the regressor and SIMCE score.

Regressor	All	C0	Q=1	Q=2	Q=3	Q=5	Q=2 & C0	Q=3 & C0
Number of answers	0.34	0.43	-	-	-	-	-	-
Proportion of coherent answers	0.42	-	-	0.33	0.28	-	-	-
Average length of answers	-	-	-	0.45	-	0.30	-	-
Average number of tokens in answers and RAE dictionary	-	-	-	0.47	-	0.28	-	-
Average number of tokens in answers with at least one digit	0.43	-	-	0.39	0.35	0.24	-	-
Average answers with at least one binary token (yes or no)	-	-	-	-	0.35	-	-	-
Sum of number of tokens in answers	0.48	-	0.31	0.42	0.35	-	0.41	0.37
Standard deviation of number of tokens in answers	0.29	-	0.16	0.33	0.22	-	0.32	0.23
Average number of tokens in answers	0.44	0.38	0.21	0.47	0.36	0.30	0.38	0.30

We will now focus on the regressors that use the linguistic attributes of the responses to open-ended questions as an estimator of the SIMCE score. For simplicity, we will report only five of these in Table 19. However, there are several other regressors that use other POS and dependency tags from student sentences. For a careful comparison, we recommend looking at Table 18 at the same time. If we consider the number of numerical modifiers (dep/nummod) tokens in coherent answers to type 3 questions (Choose and/or affirm), the correlation is 0.39. This is better than 0.30, which is achieved when only counting tokens in answers (see Table 18). The same is true if we count the number of numbers (tag/NUM) in the responses. Another important regressor is the average number of stop-words in the responses, which almost always improves the correlation instead of estimating the average

 doi:10.20944/preprints202204.0002.v1

number of words per response. For information on other linguistic attributes, please see Table 19.

Table 19. Some interesting regressors using Linguistic features of the answer. Correlation, corresponds to the Pearson coefficient between the regressor and SIMCE score.

Regressor	All	C0	Q=1	Q=2	Q=3	Q=5	Q=2 & C0	Q=3 & C0
Sum of dep/nummod token in answers	0.47	-	0.17	0.33	0.37	-	0.33	0.39
Sum of tag/NUM token in answers	0.49	-	0.21	0.38	0.36	-	0.39	0.40
Standard deviation of dep/nummod token in answers	0.35	-	0.16	0.33	0.22	-	0.32	0.23
Standard deviation of tag/NUM token in answers	0.37	-	0.06	0.31	0.30	-	0.29	0.29
Average number of stopwords in answers	0.46	0.42	0.25	0.47	0.39	0.30	0.29	0.32

Finally, the baseline model and the Open-ended model are subjected to the same evaluation criterion: two hundred and fifty 4-fold cross validations stratified by establishment. For each model, the average value of the coefficient associated with the regressor and its relevance to the model at each training stage is determined. Both the correlation with the SIMCE score and these two values are reported in one table per model (Table 20 for the baseline model and Table 21 for the Open-ended model).

First, the baseline model has two regressors with the highest weight in the model. These are (1) Grade point average for exercises, with a correlation of 0.75 with SIMCE and an average coefficient of 0.56 std SIMCE; (2) Pre-test, with a correlation of 0.70 with SIMCE and an average coefficient of 0.28 std SIMCE. The same regressors are selected by the Open-ended model as basic regressors. Historical variables such as sex and school vulnerability index are insignificant for the full model, i.e. they are not useful estimators when compared with process variables. To see other regressors , please see Table 20 and Figure 1.

The Open-ended model has 14 regressors, just like the baseline model. Of these, only 3 correspond to baseline models. These are: Grade point average for exercises, Pre-test score, and 2016 SIMCE score for the establishment. These regressors are vital as basic regressors in a linear model, given their high predictive capacity. The difference is that the other 11 regressors of the Open-ended model report other student information that can be useful for making predictions. Of the 11 selected regressors, 9 are the so-called linguistic attribute-based regressors, 6 filter for coherent responses and 5 for incoherent responses. They also filter answers by question type, where the question types chosen are types 1, 2 and 3. To see others, please see Table 21 and Figure 2.

 doi:10.20944/preprints202204.0002.v1

Table 20. Description of regressors for the Baseline model. Correlation, corresponds to the Pearson coefficient between the regressor and SIMCE score. Coefficient, average value of the coefficient associated with the regressor divided by SIMCE standard deviation (47.80 reported in [36]). Ranking, average value of the relative significance (less is better) of the modulus of the coefficient associated with the regressor for each fitted model. Both the Coefficient and the Ranking are obtained from two hundred and fifty 4-fold cross validations.

Regressor	**Correlation**	**Coefficient** (std SIMCE)	**Ranking**
Grade point average for exercises	0.75	0.57	0.0
Pre-test score	0.70	0.28	1.65
Number of words in answers	0.48	-0.11	6.11
Average number of words per response	0.48	0.29	1.87
Number of exercises	0.45	0.06	7.30
Number of exercises answered at the first attempt	0.44	0.05	8.93
Number of exercises answered more than once	0.36	0.05	8.33
2016 SIMCE score for the establishment	0.34	0.17	4.08
Number of written responses	0.34	-0.14	4.80
Exercise accuracy	-0.29	0.18	3.86
Male	0.10	0.02	11.73
Female	-0.10	-0.02	11.28
Differences between Number of exercises answered at the first attempt and that answered more than once	-0.02	-0.01	12.11
School Vulnerability Index	0.02	0.05	8.89

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 April 2022 doi:10.20944/preprints202204.0002.v1

Table 21. Description of regressors for the Open-ended model. Correlation, corresponds to the Pearson coefficient between the regressor and SIMCE score. Coefficient, average value of the coefficient associated with the regressor divided by SIMCE standard deviation (47.80 reported in [36]). Ranking, average value of the relative significance (less is better) of the modulus of the coefficient associated with the regressor for each fitted model. Both the Coefficient and the Ranking are obtained from two hundred and fifty 4-fold cross validations.

Regressor	**Correlation**	**Coefficient** (std SIMCE)	**Ranking**
Grade point average for exercises	0.75	0.49	0.0
Pre-test score	0.70	0.30	1.0
2016 SIMCE score for the establishment	0.34	0.19	2.33
Number of punctuation symbols in coherent answers to Calculate without explaining questions	0.30	0.07	8.44
Average dep/nsubj token ratio in question-dependent incoherent answers to Calculate with explaining questions	0.29	0.06	8.96
Number of coherent answers to Choice and/or affirmation questions	0.28	-0.08	7.51
Sum of dep/amod token ratio in coherent responses	0.28	0.08	7.16
Sum of dep/ROOT token ratio in question-dependent incoherent answers to Calculate with explaining questions	-0.24	-0.09	5.53
Maximum number of tag/ADV tokens in incoherent responses	0.24	0.10	4.64
Standard deviation of proportion of stopword tokens in coherent responses	-0.21	-0.07	8.37
Standard deviation of the number of tag/PRON tokens in coherent responses	0.21	-0.04	11.58
Standard deviation of number of dep/nmod tokens in question-dependent incoherent responses	0.20	-0.04	12.05
Median number of dep/case tokens in coherent responses to Calculate without explaining questions	0.20	0.05	10.06
Number of dep/obj tokens in incoherent answers to Choice and/or affirmation questions	-0.19	-0.13	3.32

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 April 2022 doi:10.20944/preprints202204.0002.v1

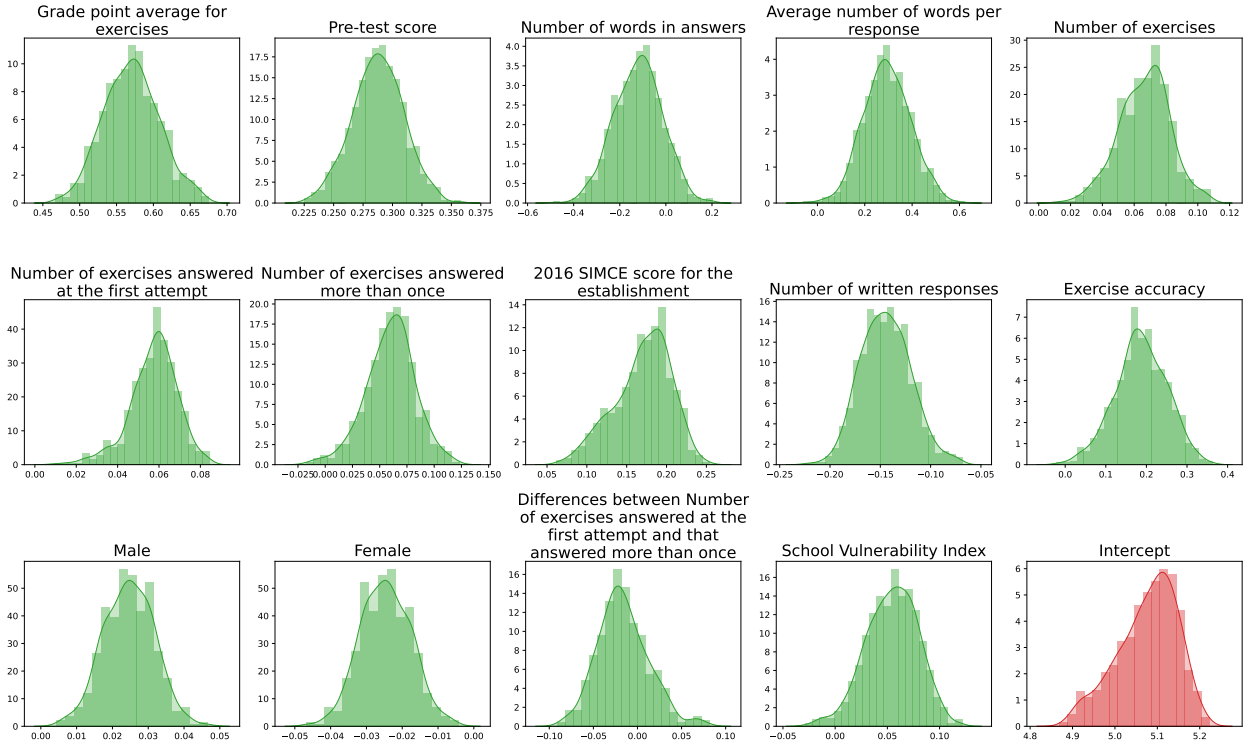

Figure 1. Distribution of each Coefficient (std SIMCE) for the Baseline model. These are obtained from two hundred and fifty 4-fold cross validations. (Green) Traditional regressors.

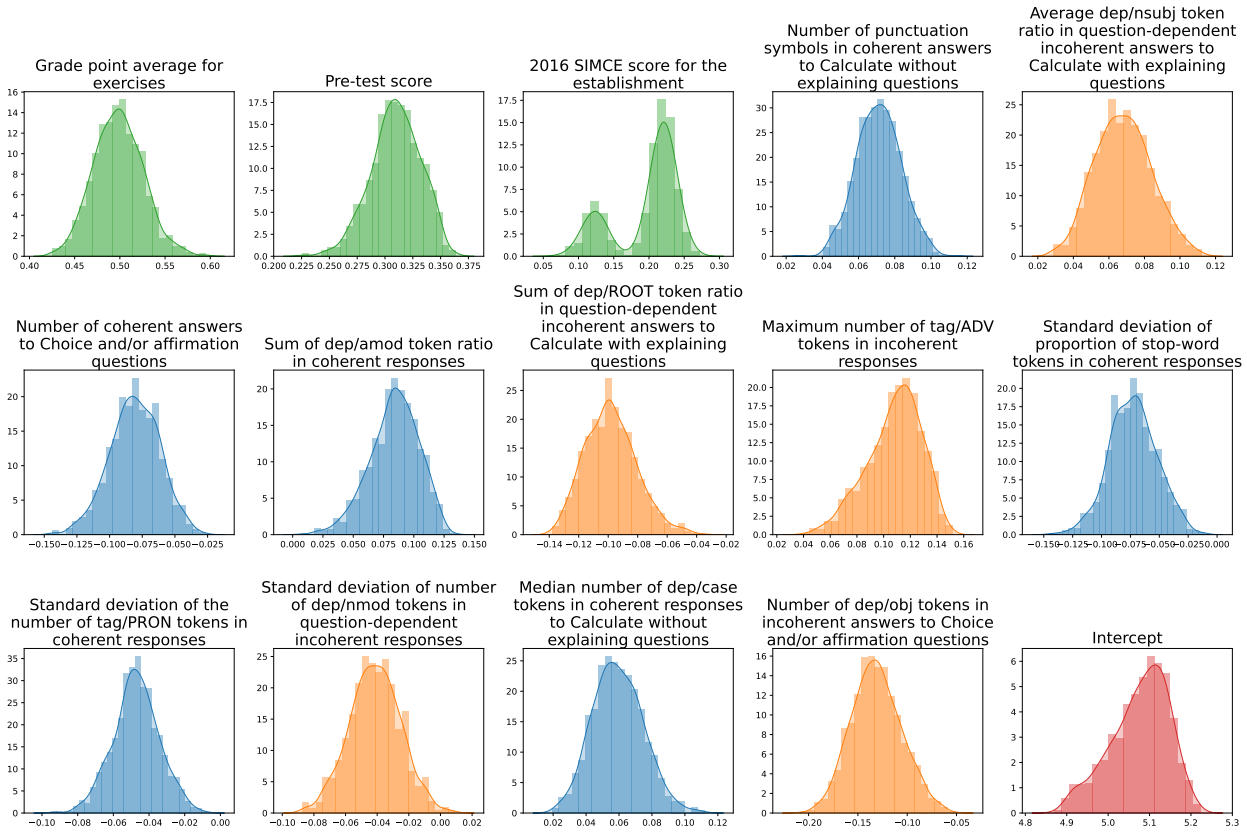

Figure 2. Distribution of each Coefficient (std SIMCE) for the Open-ended model. These are obtained from two hundred and fifty 4-fold cross validations. (Green) Traditional regressors. (Blue) Filter by coherent answers. (Orange) Filter by incoherent answers.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 April 2022 doi:10.20944/preprints202204.0002.v1

Finally, we report the correlation matrix of both models (see Figure 3) and the evaluation metrics (see Table 22). First, the correlation matrix of the Open-ended model is more homogeneous and null than the baseline model. This means that the Open-ended model regressors correlate less with each other than the baseline model regressors. This is a good feature if we want regressors with different informative estimators per student.

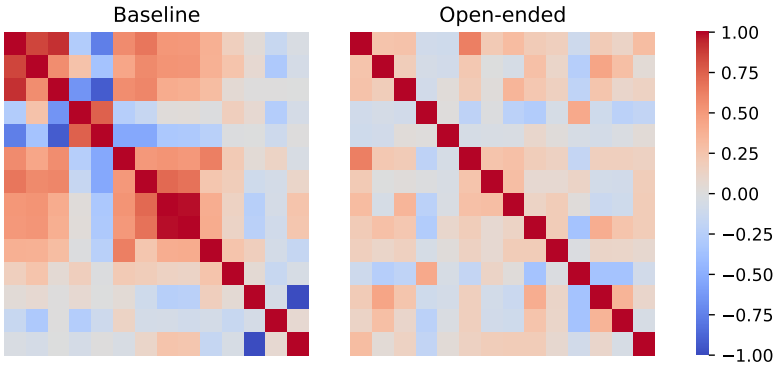

Figure 3. Correlation matrix between regressors. (**a**) Baseline model. (**b**) Open-ended model.

Next, the Open-ended model performs better in both training and testing than the baseline model, with an average R^2 of 0.70 versus 0.66 in the test set. Additionally, we consider a metric that indicates the proportion of times the open-ended model is better than the baseline model in terms of R^2. In testing, 81.1% of the time the open-ended model is better than the baseline model in terms of R^2 (similar with RMSE). See Table 22 and Figure 4 for details.

Table 22. Summary of R^2 and RMSE metrics for each model in training and test sets. These are obtained from two hundred and fifty 4-fold cross validations. RMSE is divided by SIMCE standard deviation (47.80 reported in [36]). The percentage symbol (%) correspond the proportion of times when Open-ended model is better than Baseline model, whit R^2 and RMSE respectively.

Set	R^2			**RMSE** (std SIMCE)			**Support**
	Baseline	Open-ended	%	Baseline	Open-ended	%	
Train	0.71	**0.74**	100.0	0.56	**0.53**	100.0	348.0
Test	0.66	**0.70**	81.1	0.59	**0.56**	81.1	116.0

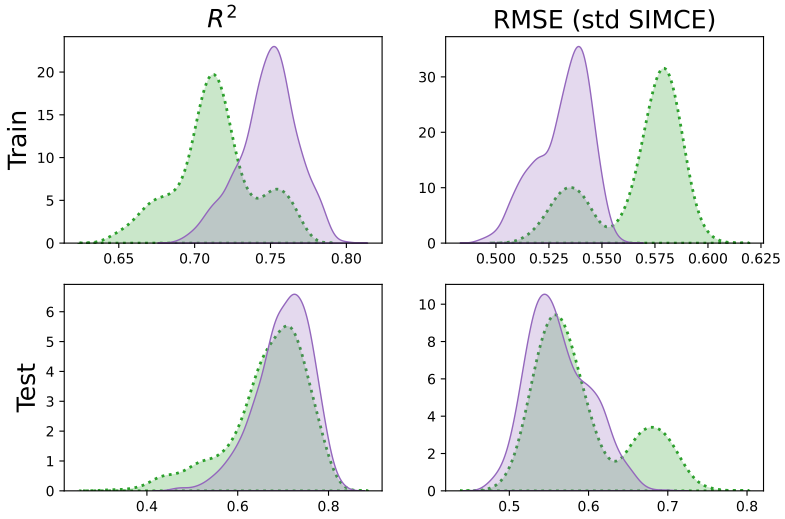

Figure 4. Distribution of R^2 and RMSE (std SIMCE) metrics for each model (in train and test sets). These are obtained from two hundred and fifty 4-fold cross validations. (Green/dotted) Baseline model. (Purple/line) Open-ended model.

 doi:10.20944/preprints202204.0002.v1

5. Discussion

5.1. Detecting coherence of answers to open-ended questions

For question type detection, the embedding-based model with BETO outperforms the linguistic attribute-based model. This may be due to the fact that the latter forgets the token order, while the former does not. This is because they are embeddings obtained from a pre-trained model with tasks that ask to predict the next token given a previous sequence.

For the detection of dependent and independent incoherence, the mixed representation of manual and BETO attributes of both question and answers outperforms the model only using attributes from BETO. This may be because manual attributes manage to condense the properties that are relevant for distinguishing incoherence in just a few layers (unlike BETO) (e.g. overlap between question and answer).

Finally, for the model for detecting the coherence of responses, the Ensemble model outperforms both the Single model and the Baselines. First, the Single model fails to distinguish between question types, where it serves as information base before estimating to better detect incoherence. For the Baselines, as well as the individual classifiers, question-answer embedding with BETO fails to generalize due to the diversity of question types. Both the Single model and the one based on BETO attributes alone are inferior to the performance of human annotators. This is not the case with the Ensemble model, which manages to improve the F1-score of Human performance by four hundredths.

One of the possible shortcomings of the Ensemble model is that if the question type detection model is wrong then most of the answers associated with that question will be detected as incoherent, given that they belong to a different question type. For example, if a type 1 question is mistaken for a type 2 question then all coherent answers will be incoherent since coherent type 2 answers are expected to have a sufficient number of words, while type 1 answers are not required to be coherent (numbers are sufficient). One solution to this is to consider probability vectors for the question type. In this way, the incoherence can be estimated through total probabilities, i.e. the probability that an answer is incoherent given that it is type 1, times the probability of being type 1, plus the probability that an answer is incoherent given that it is type 2, times the probability of being type 2, and so on.

5.2. Predicting the score on an end-of-year national standardized test

A set of estimators is constructed using the written responses. In addition, the predicted labels of question type and coherence type are used to filter the questions and obtain different regressors. Among all these regressors, the model based on responses to open-ended questions mostly only uses regressors based on linguistic attributes (9/14) and only three baseline regressors. The rest correspond to a Single indicator and a regressor based on traditional attributes from the responses. In addition, all of the regressors selected for this model distinguish between response types according to coherence and/or question type. In particular, these regressors are better estimators for predicting a final score than regressors that do not distinguish between response types. In particular, five of these regressors only use responses that are detected as incoherent, with regressors for question types 1, 2 and 3 being predominant.

The Baseline model and the model based on open-ended questions only share three of the regressors. The remaining regressors from the baseline model use simple indicators of written answers and estimators constructed using the closed-ended questions. This is not like the model based on open-ended questions, which uses the written answers to these types of questions and linguistic attributes from the answers. Finally, our results also show that the proposed model outperforms the baseline model.

We also found models based on open-ended questions with 30 regressors and an expected value of R^2 in test of nearly 0.75. We even found models with 74 regressors and an expected value of R^2 in test of nearly 0.80. These models were discarded because they had an excessive number of regressors compared to the number of regressors in the baseline model.

 doi:10.20944/preprints202204.0002.v1

6. Conclusions

To the best of our knowledge, this is the first time a study has looked at the contribution of written responses to open-ended questions on weekly formative assessments when predicting individual performance on end-of-year standardized mathematics tests by students in fourth grade.

Students only answered one or two open-ended questions each week. This is 36.4 times less than the number of answers given to closed questions. Despite this, we found that the written answers provided information that allows us to make better predictions than when only working with responses to close-ended questions. In addition to there being very few open-ended questions when compared to closed questions, the students also wrote very short answers. On average, the students wrote only 8-9 words for each answer. This is because they were not used to this type of question, let alone having to explain what they did. Even so, these written answers contain features that allow us to improve our predictions of long-term learning, as measured by scores on a national summative test. The incredible thing is that this improvement in prediction is achieved even when the summative national tests do not contain any open-ended questions. This just shows the huge potential of open-ended questions and their written responses.

In the near future, it would be important to look at the impact of an implementation with a larger number and proportion of open-ended questions. It would not be surprising if having students write more frequently and justify their responses were to have an even greater effect than that the one reported here. It also remains as future work to investigate how an early and automated warning of inconsistency could speed up the process of producing coherent explanations, as well as better explanations. On the other hand, it would be interesting to introduce instructions and explicitly teach students different strategies on how to explain their solutions to problems. We could then determine the effect this has on their adoption, as well as on the students' long-term learning. It also remains as future work to study the extent to which this process of written reasoning can be transformed from these early stages to a more rigorous form of mathematical reasoning, leading gradually to the sort of reasoning typically required in mathematical proofs.

Author Contributions: Conceptualization, R.A. and F.U.; methodology, R.A. and F.U.; software, F.U.; validation, R.A. and F.U.; formal analysis, R.A. and F.U.; investigation, R.A. and F.U.; resources, R.A..; data curation, F.U.; writing—original draft preparation, F.U.; writing—review and editing, R.A. and F.U.; visualization, F.U.; supervision, R.A.; project administration, R.A.; funding acquisition, R.A. All authors have read and agreed to the published version of the manuscript.", please turn to the CRediT taxonomy for the term explanation. Authorship must be limited to those who have contributed substantially to the work reported

Funding: This work was supported by the Chilean National Agency for Research and Development (ANID), grant number ANID/PIA/Basal Funds for Centers of Excellence FB0003..

Institutional Review Board Statement: Ethical review and approval were waived for this study, due to it being a class session during school time. The activity was revised and authorized by the respective teachers.

Informed Consent Statement: Student consent was waived due to authorization from teachers. Given that there are no patients but only students in a normal session in their schools, within school hours, and using a platform that records their responses anonymously, the teachers authorized the use of anonymized information.

Data Availability Statement: Not applicable.

Acknowledgments: Support from ANID/PIA/Basal Funds for Centers of Excellence FB0003 is gratefully acknowledged.

Conflicts of Interest: The author declares no conflicts of interest. The funder had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Mohasseb, A.; Bader-El-Den, M.; Cocea, M. Question categorization and classification using grammar based approach. *Information Processing & Management* **2018**, *54(6)*, 1228–1243.
2. Cervall, J. What the BERT?: Fine-tuning KB-BERT for Question Classification. Advanced level, Degree of Master, School of Electrical Engineering and Computer Science (EECS), 2021.
3. Bullington, J.; Endres, I.; Rahman, M. Open ended question classification using support vector machines. *MAICS 2007*, **2017**
4. Fateen, M.; Mine, T. Predicting Student Performance Using Teacher Observation Reports. In Proceedings of The 14th International Conference on Educational Data Mining; pp. 481-486. (2021)
5. Luo, J.; Sorour, E.; Goda, K.; Mine, T. Predicting Student Grade Based on Free-Style Comments Using Word2Vec and ANN by Considering Prediction Results Obtained in Consecutive Lessons. In Proceedings of the 8th International Conference on Educational Data Mining; pp. 396–399. (June 2015)
6. Canete, J.; Chaperon, G.; Fuentes, R.; Pérez, J. Spanish Pre-Trained Bert Model and Evaluation Data. PML4DC at ICLR. 2020.
7. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); Association for Computational Linguistics: Minneapolis, MN, USA, 2019; pp. 4171–4186.
8. Sepúlveda-Torres, R.; Bonet-Jover, A.; Saquete, E. "Here Are the Rules: Ignore All Rules": Automatic Contradiction Detection in Spanish. *Appl. Sci* **2021**, *11*, 3060.
9. McDermott, K. B.; Agarwal, P. K.; D'Antonio, L.; Roediger, H. L.; III, & McDaniel, M. A. Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. *Journal of Experimental Psychology* **2014**, *20*, 3–21.
10. Berninger, V. W.; Vaughan, K.; Abbott, R. D.; Begay, K.; Coleman, K. B.; Curtain, G.; Graham, S. Teaching spelling and composition alone and together: Implications for the Single view of writing. *Journal of Educational Psychology* **2002**, *94*, 291–304.
11. Hughes, E. M.; Lee, J-Y.; Cook, M. J.; Riccomini, P. J. Exploratory study of a self-regulation mathematical writing strategy. *Learning Disabilities: A Contemporary Journal* **2019**, *17(2)*, 185–203.
12. Navarro, G. A guided tour to approximate string matching. *ACM Computing Surveys* **2001**, *33(1)*, 31–88.
Wolf Garbe. 2019. Symspell. https://github. com/wolfgarbe/SymSpell.
13. Garbe, W. Symspell. Available online: https://github.com/wolfgarbe/SymSpell (March 2022).
14. Yang, Y.; Pedersen, J. A comparative study on feature selection in text categorization. *ICML* **1997**, *97*, 412–421.
15. Kowsari, K.; Jafari Meimandi, K.; Heidarysafa, M.; Mendu, S.; Barnes, L.; Brown, D. Text classification algorithms: A survey. *Information* **2019**, *10*, 150.
16. Soumya, G.K.; Shibily, J. Text classification by augmenting Bag of Words (BOW) representation with co-occurrence feature. *OSR J. Comput. Eng.* **2014**, *16*, 34–38.
17. Minaee, S.; Kalchbrenner, N.; Cambria, E.; Nikzad, N.; Chenaghlu, M.; Gao, J. Deep Learning Based Text Classification: A Comprehensive Review. *arXiv* **2020**
18. Mishra, M.; Mishra, V.K.; Sharma, H.R. Question classification using semantic, syntactic and lexical features. *International Journal of Web & Semantic Technology* **2013**, *4(3)*, p.39.
19. González-Carvajal, S.; Garrido-Merchán, E.C. Comparing BERT against traditional machine learning text classification. *arXiv* **2020**
20. Guofeng, Y.; Yong, Y. Question sentence classification of common crop disease question answering system based on Bert. *Journal of Computer Application* **2020**, *358(6)*, 1580-1586.
21. Araya, R.; Gormaz, R.; Bahamondez, M.; Aguirre, C.; Calfucura, P.; Jaure, P.; Laborda, C. ICT supported learning raises math achievement in low socioeconomic status schools. In *Design for Teaching and Learning in a Networked World*; Conole, G., Klobucar, T., Rensing, C., Konert, J., Lavoué, E., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2015; Volume 9307, pp. 383–388.
22. Araya, R.; Aljovin, E. The effect of teacher questions on elementary school students' written responses on an online STEM platform. *Advances in Human Factors in Training, Education, and Learning Sciences. Advances in Intelligent Systems and Computing* **2017**, 596, 372-382. Springer.
23. Araya R., Jiménez A., Aguirre C. Context-Based Personalized Predictors of the Length of Written Responses to Open-Ended Questions of Elementary School Students. In: Sieminski A., Kozierkiewicz A., Nunez M., Ha Q. (eds) *Modern Approaches for Intelligent Information and Database Systems*. Studies in Computational Intelligence, 2018, vol 769. pp 135-146, Springer, Cham
24. Araya, R. Teacher Training, Mentoring or Performance Support Systems. In *AHFE 2018: Advances in Human Factors in Training, Education, and Learning Sciences*; Nazir, S., Teperi, A.M., Polak-Sopínska, A., Eds.;Advances in Intelligent Systems and Computing; Springer International Publishing: Cham, Switzerland,2019; Volume 785, pp. 306–315.
25. Araya, R.; Diaz, K. Implementing Government Elementary Math Exercises Online: Positive Effects Found in RCT under Social Turmoil in Chile. *Educ. Sci.* **2020**, 10, 244.
26. Bjork, R.A.; Bjork, E.L. Desirable difficulties in theory and practice. *J. Appl. Res. Mem. Cogn.* **2020**, 9, 475–479
27. Black, P.; Wiliam, D. Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice* **1998**, 5(1), 7-74.
28. Dunlosky, J.; Rawson, K.A.; Marsh, E.J.; Nathan, M.J.; Willingham, D.T. Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychol. Sci. Public Interest* **2013**, 14, 4–58

29. Gezer, T.; Wang, C.; Polly, A.; Martind, C.; Pugaleee, D.; Lambert, R. The Relationship between Formative Assessment and Summative Assessment in Primary Grade Mathematics Classrooms. *International Electronic Journal of Elementary Education* **2021**, 13(5), 673-685.
30. Hodgen, J.; Foster, C.; Marks, R.; Brown, M. Improving Mathematics in Key Stages Two and Three: Evidence Review. London: Education Endowment Foundation. Available at: https://educationendowmentfoundation.org.uk/tools/guidance-reports/maths-ks-two-three/ (2017)
31. McDaniel, M. A.; Little, J. L. Multiple-choice and short-answer quizzing on equal footing in the classroom: Potential indirect effects of testing. In Dunlosky, J., Rawson, K. A. (Eds.), The Cambridge handbook of cognition and education (p. 480–499). Cambridge University Press. 2019
32. Rittle-Johnson, B.; Loehr, A.; Durkin, K. Promoting Self-Explanation to Improve Mathematics Learning: A Meta-Analysis and Instructional Design Principles. *ZDM Mathematics Education* **2017**, 49(4), 599-611
33. Soderstrom, N.; Bjork, R. Learning Versus Performance: An Integrative Review. *Perspectives on Psychological Science* **2015**, 10(2), 176–199
34. Wang, T.; Ma, F.; Wang, Y.; Tang, T.; Zhan, L.; Gao, J. Towards Learning Outcome Prediction via Modeling Question Explanations and Student Responses. Proceedings of the 2021 SIAM International Conference on Data Mining (SDM). pp. 693 - 701
35. Zheng, G.; Fancsali, S.; Ritter, S.; Berman, S. Using Instruction-Embedded Formative Assessment to Predict State Summative Test Scores and Achievement Levels in Mathematics. *Journal of Learning Analytics* **2019**, 6(2), 153 —174.
36. Ulloa Miranda, O. A. Estimación de desempeño en evaluación sumativa, con base en evaluaciones formativas usando modelos espacio estado. Thesis to obtain the degree of Master in Applied Mathematics. (Universidad de Chile, 2021)
37. Leardi, R.; Boggia, R.; Terrile M. Genetic algorithms as strategy for feature selection. *J. Chemometrics* **1992**, 6, 267.
38. Araya, R.; Arias Ortiz, E.; Bottan, N.; Cristia, J. Does Gamification in Education Work? Experimental Evidence from Chile. Available online: https://publications.iadb.org/publications/english/document/Does_Gamification_in_Education_Work_Experimental_Evidence_from_Chile_en_en.pdf (accessed on 23 June 2020).
39. Bicer, A.; Capraro, R.M.; Capraro, M.M. Integrating writing into mathematics classroom to increase students' problem solving skills *International Online Journal of Educational Sciences* **2013**, 5(2), 361-369.
40. Anozie, N.; Junker, B. W. Predicting end-of-year accountability assessment scores from monthly student records in an online tutoring system. Paper presented at the AAAI workshop on educational data mining, Menlo Park, CA. 2006.
41. Nguyen, D.; Smith, N.A.; Rose, C. Author age prediction from text using linear regression. In Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, Portland, OR, USA, 24 June 2011; pp. 115–123.
42. Schwartz, R.; Sap, M.; Konstas, I.; Zilles, L.; Choi, Y.; Smith, N. A. (2017). The effect of different writing tasks on linguistic style: A case study of the ROC story cloze task. *arXiv* **2017**
43. Pugalee, D. K. Writing, mathematics, and meta-cognition: Looking for connections through students' work in mathematical problem solving. *School Science and Mathematics* **2001**, 101(5), 236-245.
44. Martin, C. L. Writing as a tool to demonstrate mathematical understanding. *School Science and Mathematics* **2015**, 115(6), 302–313.
45. Powell, A.B. Capturing, examining and responding to mathematical thinking through writing. *The Clearing House* **1997**, 71(1), 21–25.
46. Yamaç, A.; Öztürk, E.; Mutlu, N. Effect of digital writing instruction with tablets on primary school students' writing performance and writing knowledge. *Computers & Education* **2020**, 157, p. 103981
47. Berninger, V.; Vaughan, K.; Abbott, R.; Begay, K.; Byrd, K.; Curtin, G.; Hawkins, J. M.; Graham, S. Teaching spelling and composition alone and together: Implications for the simple view of writing. *Journal of Educational Psychology* **2002**, 94, 291-304
48. Namoun, A.; Alshanqiti, A. Predicting Student Performance Using Data Mining and Learning Analytics Techniques: A Systematic Literature Review. *Appl. Sci.* **2021**, 11, 237.
49. Yamasari, Y.; Rochmawati, N.; Putra, R. E.; Qoiriah, A.; Yustanti, W. Predicting the Students Performance using Regularization-based Linear Regression. In 2021 Fourth International Conference on Vocational Education and Electrical Engineering (ICVEE), IEEE 2021, pp. 1-5.
50. Kaye, L.K.; Malone, S.A.; Wall, H.J. Emojis: Insights, affordances, and possibilities for psychological science. *Trends Cogn. Sci.* **2017**, 21, 66–68