International Journal of Artificial Intelligence in Education
https://doi.org/10.1007/s40593-021-00283-x

ARTICLE

Machine Learning and Hebrew NLP for Automated Assessment of Open-Ended Questions in Biology

Moriah Ariely[1] · Tanya Nazaretsky[1] · Giora Alexandron[1]

Accepted: 19 September 2021Published online: 03 January 2022

Abstract

Machine learning algorithms that automatically score scientific explanations can be used to measure students' conceptual understanding, identify gaps in their reasoning, and provide them with timely and individualized feedback. This paper presents the results of a study that uses Hebrew NLP to automatically score student explanations in Biology according to fine-grained analytic grading rubrics that were developed for formative assessment. The experimental results show that our algorithms achieve a high-level of agreement with human experts, on par with previous work on automated assessment of scientific explanations in English, and that ∼500 examples are typically enough to build reliable scoring models. The main contribution is twofold. First, we present a conceptual framework for constructing analytic grading rubrics for scientific explanations, which are composed of dichotomous categories that generalize across items. These categories are designed to support automated guidance, but can also be used to provide a composite score. Second, we apply this approach in a new context – Hebrew, which belongs to a group of languages known as Morphologically-Rich. In languages of this group, among them also Arabic and Turkish, each input token may consist of multiple lexical and functional units, making them particularly challenging for NLP. This is the first study on automatic assessment of scientific explanations (and more generally, of open-ended questions) in Hebrew, and among the firsts to do so in Morphologically-Rich Languages.

Keywords Automatic scoring · Scientific explanations · Natural language processing · Morphologically rich languages · Biology education

Moriah Ariely and Tanya Nazaretsky contributed equally to the paper.

✉ Moriah Ariely
moriah.ariely@weizmann.ac.il

Extended author information available on the last page of the article.

Introduction

Enhancing students' ability to formulate scientific explanations is considered a major goal of science education (National Research Council (NRC), 2012). Providing explanations requires skills and performances such as task framing, weighing the value and relevance of information, and assembling disparate knowledge elements into clear, logical, coherent explanatory structures (Nehm et al., 2012; Berland & Reiser, 2009; Osborne & Patterson, 2011). However, research has shown that a considerable number of students are unable to construct proper explanations (McNeill & Krajcik, 2008; Nehm et al., 2012; Songer & Gotwals, 2012; Tang, 2016).

The feedback that students receive through instruction is a key factor in developing their scientific writing skills, as it is the primary means through which they can evaluate and improve their writing (Allen et al., 2016; Roscoe et al., 2013; Hattie & Timperley, 2007). Students' responses to open-ended questions provide opportunities for assessing their abilities to construct and communicate valid scientific explanations. However, proper assessment of open-ended questions is a costly and time-consuming process, which typically requires developing grading rubrics and applying them to student responses (Li et al., 2017; Nehm et al., 2012). This process is also subject to rater variability and reliability threats, as well as to subgroup biases (Bridgeman et al., 2012; Madnani et al., 2017b), potentially resulting in inaccurate human scoring (Li et al., 2017; Nehm et al., 2012). Faced with large class sizes, teachers are struggling to find the time to devote to this process, resulting in a considerable delay in the feedback students receive, and affecting its quality (Allen et al., 2016). Consequently, revising written explanations based on personalized guidance rarely occurs in science classrooms (Tansomboon et al., 2017). This is consistent with statistics computed on the online learning platform that is developed within the Department of Science Teaching at Weizmann Institute, which serves more than a thousand high-school science teachers in Israel. A database search yielded that less than 10% of the open-ended items administered to students are eventually graded.

Technology holds much promise for improving this process. In recent years there have been efforts to develop computer-based systems for writing assessment and instruction, which are capable of addressing many of the aforementioned disadvantages (e.g., Allen et al., 2016; Klebanov et al., 2017; Gerard & Linn, 2016; Li et al., 2017; Nehm et al., 2012; Strobl et al., 2019; Tansomboon et al., 2017; Wilson et al., 2017; Zhai et al., 2020; Maestrales et al., 2021). Automated scoring techniques can provide students with automated feedback as they write scientific explanations or immediately following the writing task (Li et al., 2017), and they have other advantages for assessment, such as objectivity, standardization, and efficiency (Cohen et al., 2018).

Automated formative assessment of constructed-response items in science education is typically based on invoking feedback that aims to close the gap between the expected and observed performance, as identified using content-based scoring rubrics (Liu et al., 2014; Tansomboon et al., 2017; Nehm et al., 2012). Rubrics can be holistic or analytic (Jescovitch et al., 2019b). An advantage of analytic bins is that they can be

directly connected to feedback (Rahimi et al., 2017). Another advantage is that they can be combined to provide a composite score that may be more accurate than scores based on holistic rubrics, especially on complex items (Jescovitch et al., 2021).

Our Research We study methods for automatic assessment of scientific explanations, with the goal of supporting formative feedback. We follow the aforementioned approach and use fine-grained analytic rubrics, which are designed to support automated guidance, and are amenable to machine-learning modeling. The rubrics are based on the idea of decomposing a correct explanation into a collection of binary categories (analytic bins), each representing a property that should be present in a correct response, and may be associated with corresponding feedback in case that it is missing, partial, or wrong. Our machine learning modeling approach is to build a scoring model (classifier) per category, whose output can be used to determine whether to invoke the feedback or not (according to some pedagogic policy that may give preference to certain types of error). In addition, the output of all the classifiers can be combined to form a holistic composite score. The rubrics are described in detail in the next sections. In addition to their advantage for feedback generation, and for unpacking complex knowledge, we demonstrate that machine learning models built on analytic bins that capture cross-item concepts can effectively *generalize between items* that require the knowledge categories modeled by these bins.

We study this topic in a language in which NLP-based automatic assessment of open-ended questions was not studied before – Hebrew. Hebrew is a medium-sized, low-resource language that belongs to a group of languages that is referred to in the NLP research community as *Morphologically-Rich Languages* (MRLs) (Tsarfaty et al., 2013). (We extend on the special challenges that Hebrew, and MRLs in general, pose to NLP, and on the (scarce) research that was conducted on the applications of Hebrew NLP to education, in the Literature Review.)

Previous research noted that NLP pipelines built with English in mind may not work well for MRLs (Tsarfaty et al., 2020). As this is the first study on NLP-based automatic assessment in Hebrew, we start with a case study – automated assessment of an instrument containing three constructed-response items on cellular respiration. All items refer to the same biological mechanism, but the item context and surface features vary. (We elaborate on the instrument in "The Instrument" Subsection.) The rationale that guides our research is that we first wish to study whether we can build machine learning models that are capable of performing an in-depth assessment of content knowledge, on a level that is useful for high-school Biology teachers and is required in order to persuade them to adopt such tools, before trying to scale to additional topics and instruments.

Following this, the concrete objectives of this study are to evaluate whether Hebrew NLP-based scoring models can accurately grade student responses to cellular respiration constructed-response items according to fine-grained analytic rubrics, on par with human experts. Once this proof-of-concept is established, we will feel confident to use this pipeline as the basis for automated intervention in real classrooms with this topic, while scaling to additional topics in Biology.

These objectives are formalized into the following research questions (RQs):

1. **RQ1**: Can our scoring models accurately grade unseen responses to a certain item from the instrument after being trained on graded responses of the *same* item (within-item transfer)?
2. **RQ2**: Can our scoring models accurately grade unseen responses to a certain item after being trained on graded responses to *different* items from the instrument (between-item transfer)?
3. **RQ3**: Can our scoring models accurately grade unseen responses to a certain item after being trained on graded responses to all the items of the instrument (instrument-level modeling)?
4. **RQ4**: How many human-scored responses are needed to build scoring models that reach a satisfactory level of agreement with human experts (e.g., are 200 graded responses enough)?

The rationale behind these RQs is as follows. First, we wished to explore if our algorithms can reach a satisfactory level of accuracy (interpreted as a high level of inter-rater agreement with human experts) in the most narrow scenario – being trained and applied on responses to the same item. However, collecting a large number of labeled examples is not always feasible, especially in formative assessment applications in which the content is frequently updated. Thus, we explored possible ways to mitigate this 'cold start' problem and expedite the process of adding models for new items by generalizing existing grading models to new items, or by pooling data from several items from the same instrument. Under the same rationale of investigating the performance under realistic constraints, we studied the amount of data that are actually required for reaching a high level of accuracy.

We evaluated these RQs on data of ∼700 high-school students who responded to the instrument as part of their study of the topic of "The Respiratory and Blood Systems". Student responses were graded according to the grading rubrics mentioned above, and the labeled dataset was used to train and evaluate Hebrew NLP-based machine-learning scoring models.

Contribution First, this study extends the body of research on the applications of NLP to automatic assessment in non-English contexts. The results of our case study demonstrate that automatic scoring models for scientific explanations in Hebrew can achieve a high level of agreement with human experts. This is the first study on NLP-based automatic assessment of scientific explanations (and more generally, of open-ended questions) in Hebrew – a medium-sized, low-resource language – and among the firsts to study this in the context of MRLs (e.g., Hebrew, Arabic, Turkish). These languages pose considerable challenges for NLP, and are under-represented in the research on applications of NLP to educational technologies, making the problem both scientifically interesting and educationally important.

Second, this study extends our understanding of the pros and cons of analytic grading rubrics for NLP-based automatic assessment. We demonstrate that fine-grained analytic grading rubrics that are designed for formative assessment purposes and aim

to capture conceptual understanding, can be reliably learned by deep learning models and that models built on cross-item analytic components are highly generalizable between items that share these components.

Third, our results on the amount of training data that is required to build reliable machine learning-based scoring models reinforce previous studies that reported that ∼500 labeled examples are typically sufficient to achieve a high level of agreement with human experts.

The rest of the paper is organized as follows. First, we survey relevant research. Second, we describe the instrument and the Grading Rubric. Third, we describe the Data Collection and Human Grading process. Fourth, we describe the NLP and Machine Learning Modeling. We then present the Results, followed by a Discussion and Summary.

Literature Review

Machine Learning for Automated Assessment in Science Education

Assessment may be for *formative* or *summative* purposes. Summative assessment is aimed at evaluating learning and instruction outcomes, while formative assessment is aimed at collecting feedback to support teachers' instructional decisions and students' learning (Zhai, 2021). These are not separate or fixed paradigms (Taras, 2005), and assessment may serve both purposes simultaneously (Alexandron et al., 2020). However, assessment becomes 'formative' when evidence is actually used to adapt the instruction to meet student needs (Black & Wiliam, 1998).

The outcomes of machine learning (ML)-based assessments can be used to provide immediate feedback to teachers and students, and thus, ML is increasingly used in web-based inquiry, game-based assessment, simulation assessment, and adaptive learning (Zhai, 2021). Application of ML in science assessment primarily involves text recognition and automatic scoring of constructed-response items (Zhai et al., 2020). New technologies for writing assessment are based on Natural Language Processing (NLP) tools. Advances in this field, together with the availability of unprecedented amounts of data from digital learning environments, have led to an increasing interest in using NLP to address the needs of teachers and students (Litman, 2016).

Research on automated assessment has demonstrated that ML-involved science assessments have many significant advantages compared with traditional assessments (Zhai, 2021). For example, automated scoring systems were found to reduce the time required for grading literacy assignments and to have a positive effect on the quantity of feedback (Matthews et al., 2012), and they were also found to be a powerful and cost-effective tool for assessing students' written explanations of evolutionary change (Nehm et al., 2012). Maestrales et al. (2021) examined ML's capacity to automatically score multi-dimensional science assessments tasks, showing that ML algorithms can successfully classify student responses, potentially facilitating

the transition to multi-dimensional assessment that can be done more quickly than when employing human scoring alone. Indeed, student-facing automated scoring systems designed for formative assessment on constructing scientific explanations were shown to drive significant learning gains (Gerard & Linn, 2016; Tansomboon et al., 2017; Allen et al., 2016; Graesser et al., 2014; McNamara et al., 2013). Such tools can also serve *teachers*, for example by providing them with real-time information on conceptual gaps, which can be used to assign adaptive guidance (Roschelle et al., 2013; Tansomboon et al., 2017). By that, automated scoring tools can strengthen the teacher's role (Gerard & Linn, 2016). More in the context of constructing scientific arguments, Zhu et al. (2017), reported that the automated feedback supported students in providing scientific arguments. A recent study strengthens this connection between revisions enabled by the automated formative system, and students' performance and learning gains (Zhu et al., 2020).

Although research on ML-based assessments has already provided encouraging results on applications of NLP for automated assessment in science education, most of the applications of this technology are still employed for summative purposes (Woods et al., 2017). In addition, only a limited number of studies focused on K-12 science learning (Zhai et al., 2020), and there is clearly a need for a multi-lingual perspective (Klebanov & Madnani, 2020). Our research aims for these gaps.

Automated Content Scoring of Constructed-response Items

Within the context of automated assessment in science education, we focus on content scoring of constructed-response items. Responses to such items are typically short in length (roughly between one phrase and one paragraph) (Burrows et al., 2015), and thus are also referred to as Automated Short Answer Scoring (ASAS).

ASAS aims to assess what the student knows, has learned, or can do in a specific subject area, thus the focus is on the content rather than on the writing style, spelling mistakes, or grammatical errors (Burrows et al., 2015; Madnani et al., 2017a). Assessing the content requires attention to whether students are using the correct concepts, accurately describing the relationships among them, and providing an adequate level of detail (Madnani et al., 2017a). ASAS is therefore a challenging application of NLP, and it typically relies on response data that are manually labeled according to evaluation rubrics. Previous studies of content scoring provided valuable insights and implications of different aspects which impact the scoring performance such as training sample size, the minimum number of examples per score level, the length of the responses (Heilman & Madnani, 2015), data quality (Yao et al., 2020), and also of different linguistic features that are highly predictive for the automated scoring (Padó, 2016).

Several NLP tools, and their application to ASAS in science education, have been proposed. These include applying the Summarization Integrated Development Environment (SIDE) (Mayfield & Rosé, 2013) for scoring explanations in Biology (Nehm et al., 2012; Ha et al., 2011), using SPSS Text Analysis (SPSSTA) to identify terms throughout the text (Weston et al., 2013; Nehm & Haertig, 2012), EvoGrader that automatically scores constructed explanations using ML algorithms (Moharreri et al.,

2014), and using ETS' c-Rater to score scientific explanations based on the presence of central concepts (Liu et al., 2014). Although these studies showed a good human-computer agreement, each application has its disadvantages as well (Li et al., 2017). For example, SIDE is limited to identifying the presence of concepts within students' responses and is, therefore, not useful at scoring students' competencies at reasoning, and SPSSTA was unable to automatically produce ML algorithms from a trained data (Ha et al., 2011). EvoGrader succeeded in producing human-like scoring of key evolutionary concepts but needs retraining in order to generalize to other domains (Moharreri et al., 2014), and c-Rater sensitivity to variations in the phrasing of central concepts needed to be improved (Liu et al., 2016).

There are two main machine-learning strategies for ASAS: Reference-based and response-based (Madnani et al., 2017a). In the reference-based approach, student responses are scored based on their similarity to reference answers provided by experts, or by selecting high-scoring responses (e.g., Horbach et al., 2013; Pado & Kiefer, 2015). In the response-based approach, human experts provide labeled data that are used to train machine learning models (e.g., Zesch et al., 2015; Zhu et al., 2016). According to Sakaguchi et al. (2015), if sufficient human-scored data is available, the response-based approach is superior to the reference-based one in terms of grading accuracy. However, its main drawback is that developing high quality rubrics, and applying them to response data, is a highly human resource intensive process, especially that it is typically required to build separate scoring models for each item (Madnani et al., 2017a).

As noted, the supervised approach to ASAS is based on response data that are labeled according to grading rubrics, which can be either *holistic* or *analytic*. In holistic scoring, the rater makes an overall judgment about the quality of the response (Jonsson & Svingby, 2007), and the rubric typically consists of a description of student performance corresponding to different levels of proficiency (Wang et al., 2021). As opposed to holistic scoring, in analytic scoring, the rater assigns a score to different dimensions being assessed in the task (Jonsson & Svingby, 2007). The analytic bins can be directly connected to feedback (Rahimi et al., 2017), making this approach especially useful for helping teachers and students to identify strengths and learning needs, namely, for formative assessment (Wang et al., 2021). In addition, a composite score can be computed by summing over the analytic bins. It was recently shown that composite analytic scoring may achieve more accurate results than holistic scoring (Jescovitch et al., 2021; Wang et al., 2021), especially in more complex items (Jescovitch et al., 2021). The analytic approach can reduce coding complexity and improve the model development process (Jescovitch et al., 2019a), and it was also found to be more reliable in assessing content components of reasoning and provide specific feedback to the student (Jescovitch et al., 2019a; Yune et al., 2018). However, developing analytic grading rubrics, and applying them to response data, is more time-consuming compared to the holistic scoring approach (Jescovitch et al., 2021). As we demonstrate in this study, machine learning models built on analytic bins that capture conceptual knowledge may generalize well across constructed-response items, underlining a potentially important advantage of their modularity to ASAS in science education.

Applications of Hebrew NLP to Education

NLP applications are built over fundamental NLP tools that conduct the basic computational processing tasks (e.g., Parsing). In Hebrew, until recently, the lack of open Hebrew NLP (HNLP) resources limited the ability to develop HNLP applications. As a result, research on applications of NLP to science education in Hebrew is scarce (Ariely et al., 2020). To date, we are familiar with only two publications in this domain – the work of (Segal et al., 2017), which developed an NLP-based tool to support teachers in managing group work, and our preliminary work on formative assessment of scientific writing in Biology (Ariely et al., 2020). (Related work on Automated Essay Scoring is conducted at the National Institute for Testing and Evaluation (Cohen & Ben-Simon, 2011; Cohen et al., 2018).)

Hebrew belongs to a group of languages that are referred to in the NLP research community as Morphologically-Rich Languages (MRLs). In MRLs each input token may consist of multiple lexical and functional units, which makes MRLs particularly challenging for NLP (Tsarfaty et al., 2013; Tsarfaty et al., 2019). Other examples of MRLs are Arabic, which is closely related to Hebrew (the two are Semitic languages), and Turkish, both with a much larger speaker population than Hebrew. Similarly to the situation in Hebrew, the body of work on NLP-based automatic grading of scientific writing in Turkish and Arabic (and other MRLs) is very limited. For example, (Çınar et al., 2020) is the first study on NLP-based automatic assessment of scientific explanations in Turkish (the context is different than ours – Physics in Higher Education, while we work in Biology in K-12). Concerning Arabic, the recent review of (Flor & Cahill, 2020) on automated scoring of open-ended questions actually mentions only Automated Essay Scoring in this language. Gomaa and Fahmy (2014) translated responses to open-ended questions to English to overcome the lack of NLP resources in Arabic. We are unfamiliar with more recent research on NLP-based scoring of open-ended questions in Arabic.

Fortunately, in the last years there is a significant growth in the availability of fundamental NLP tools for MRLs, and specifically, for Hebrew (Sheinfux et al., 2015; Jacobs et al., 2020; Seddah et al., 2013; Tsarfaty et al., 2019). Our research builds on these advances to pioneer research and development of NLP applications to science education in Hebrew, centering on the formative assessment of scientific explanations.

The Instrument and Grading Rubric

The Instrument

The instrument consisted of three open-ended questions (items) about the effect of smoking, anemia, and high altitude on physical activity as presented below (hereafter referred to as Smoking, Anemia, and Height items). All items deal with the influence of blood oxygen levels on humans' ability to do a physical activity, and should include references to relevant information such as the role of red blood cells and Hemoglobin, blood circulation, and cellular respiration. These topics are generally

taught in 10th grade, in the context of the respiratory and circulatory systems, which are taught as part of the syllabus core subject "The human body" (Israeli Ministry of Education, 2011).

The items chosen for the instrument are typical open-ended questions in Biology. Different versions of these questions appear frequently in a variety of teaching materials and exams, including in the Israeli Matriculation Exams in Biology in recent years. The instrument contained the following items (translated from Hebrew):

1. **Smoking:** The smoke from cigarettes contains several harmful substances, including the gas carbon-monoxide (CO). CO is released from cigarettes while smoking, and has a stronger tendency than oxygen to bind to Hemoglobin. Explain how high levels of CO make it difficult for smokers to exercise.
2. **Anemia:** A person was found to have low levels of red blood cells in his blood test (anemia). This person complained to his doctor about weakness and difficulty to exercise. Explain how low levels of red blood cells make it difficult for people with anemia to exercise.
3. **Height:** In high altitudes the air is thin (low atmospheric pressure) and has low amounts of oxygen. In a study, the achievements of two groups of athletes, sent to a running competition in Mexico (located about 2000 meters above sea level), were compared. Prior to the competition, both groups had similar achievements and they received the same conditions (food, training, and rest). Group A traveled to Mexico a long time before the competition (about a month before), while group B traveled to Mexico just a few days before it. It was found that the achievements of group A in the competition were higher than the achievements of group B. Explain how the differences in the time each group spent in Mexico prior to the competition resulted in differences in the groups' achievements during the competition.

The Grading Rubric

Overall rationale The majority of explanations in the science classroom attempt to provide an account that specifies what happened and why it occurred, namely, constructing causal accounts of phenomena (Berland & Reiser, 2009; Osborne & Patterson, 2011). Causal explanations are at the center of biology education. Such explanations should go beyond defining or describing a specific process, and they require a chain of reasoning for the phenomenon to be explained (Kampourakis & Neibert, 2018). Accordingly, the rubrics we developed include a set of categories, each representing an important element in the causal chain, which is needed for the explanation to be accurate and complete. An example of the causal chain for the Smoking item is presented in Fig. 1.

As shown in the example, the chain of reasoning includes all the main key events and components ('steps') that are relevant to the underlying phenomenon to be explained (represented by the dark grey boxes in Fig. 1). The causal relations between 'steps' (represented by the light grey boxes) are also important, because they provide some information about *how* one step leads to the other, and not only what those steps are. Therefore, filling the causal relations suggests that one knows more about the

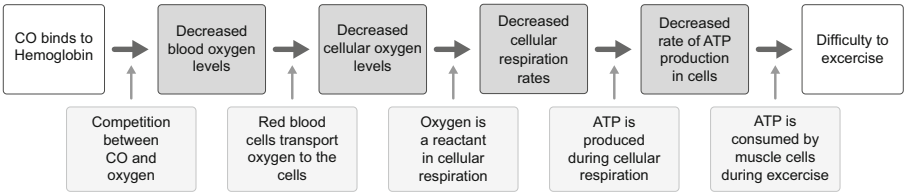

Fig. 1 An example of the expected causal chain of reasoning for the Smoking item. The information in the first and last boxes (white fill) was given in the question. Grey boxes refer to the causal chain, including key events (dark grey) and causal relation (light grey)

causal process than merely saying that one event causes or affects the other (Ross, 2020). Note that 'steps' and causal relations in the expected causal chain were used as categories in the scoring rubrics for each item.

The cornerstone of our approach is that the grading rubric decomposes a correct explanation into a collection of binary categories, each representing an essential property that the student explanation should include, and may be associated with appropriate guidance to the student in case that this property is not met. (We note that reporting results from a formative assessment intervention study is beyond the scope of the current paper.) The categories are scored independently, but these scores can be combined to form a composite score for each item, so this modeling can also serve holistic grading.

Our approach allows separate bins to be generalized to other items having these bins *as part* of their chain of reasoning, even if the causal chain is not entirely the same. This is in contrast to grading rubrics that are typically item-specific (Madnani et al., 2017a), and may reduce the time and effort of developing grading rubrics for new items. To demonstrate this, we have chosen representative items from a central topic in biology education, which frequently appears in various formats and contexts, making this case study on analytic grading rubrics generalizable to a large set of items. In addition, this methodology provides a conceptual approach for developing scoring rubrics for causal explanations – an important practice in biology education – which centers on the underlying chain of reasoning.

The Rubrics In all three items, the expected explanations include chains of reasoning that are composed of two main parts. The first part of the causal chain leads to changes in cellular oxygen levels due to CO release (Smoking item), low levels of red blood cells (Anemia item), or high altitude (Height item). Thus, the first part of the explanations' causal chain is different in each of the three questions, but in all three explanations, it leads to changes in cellular oxygen levels. The second part of the causal chain leads to changes in the ability to exercise due to the changes in cellular oxygen levels, and it is similar in all three explanations. It should be noted, that while the Smoking and Anemia items are similar in their complexity, and require a relatively simple chain of reasoning, the Height item is more complex since it requires a comparison between two groups, and it has more than one optional chain of reasoning that provides an explanation of the phenomenon.

We created a grading rubric for each of the three items in the instrument. We took a top-down approach for developing the rubric, namely, the causal explanation and the underlying reasoning chains of cause and effect were defined by an expert prior to data collection. The rubric development had two main phases. First, a basic rubric was initially developed by the first author (a domain expert). The grading rubric was further refined after grading a subset of items through an iterative process with a second domain expert until 100% agreement was reached on the basic items of the grading rubric. Second, the rubric was further developed, revised, and refined with the help of five experienced biology teachers, as follows: We provided the teachers with basic information about AI and automated feedback, and then the instrument was presented to them. We asked the teachers to provide categories for the correct chain of reasoning for each of the items in the instrument. The teachers provided the chain of reasoning for each item and built the rubric categories accordingly for all the items. Then, we compared the expert categories with the categories composed by the teachers. The initial agreement was over 90% for all items. We discussed with the teachers the causal chain of reasoning and the categories until 100% agreement

Table 1 The categories of the grading rubric. The 'relevant' column indicates which rubric categories apply to each item. The '%correct' reflects the difficulty of the category, measured as the percentage of students who got the category correct

		Anemia Item		Smoking Item		Height Item	
		relevant	% correct	relevant	% correct	relevant	% correct
A	Reference to both groups	–	–	–	–	+	60%
B	Changes in the amount of red blood cells	–	–	–	–	+	47%
C	Changes in oxygen levels that bind to Hemoglobin/ red blood cells	–	–	+	79%	+	31%
D	The role of Hemoglobin/ red blood cells in oxygen transportation	+	60%	+	26%	+	8%
E	Changes in oxygen levels in the body (general)	+	77%	+	71%	+	36%
F	Changes in oxygen levels in the cells (micro level)	+	51%	+	51%	+	21%
G	Oxygen is a reactant in cellular respiration/energy production	+	38%	+	33%	+	13%
H	Changes in cellular respiration rate	+	33%	+	31%	+	16%
I	Using the term cellular respiration	+	54%	+	51%	+	26%
J	Energy/ATP is produced during cellular respiration	+	27%	+	19%	+	9%
K	Changes in energy/ATP levels	+	54%	+	48%	+	23%
L	Using the term energy/ATP	+	65%	+	57%	+	30%
M	Energy is consumed during exercise	+	25%	+	29%	+	8%

was reached among all members of the group. The final rubrics (for the three items) include the agreed categories and are presented in Table 1.

Data Collection and Human Grading

The research population for this study is high school students. The research sample included 669 students, equally distributed in grades 10-12. Gender distribution was 70% females (representative of the gender distribution among high school Biology students). Students were from approximately 25 high schools of varied geographic and socio-economic status (based on the school location).

The instrument was distributed to the teachers via teachers' professional communities, and teachers self-selected whether they wish to administer it to their students. Due to ethical regulations, answers were collected anonymously, and students were requested to fill in only their grade, gender, and geographic location. The teachers were instructed to administer the instrument to the students after teaching the relevant topics within the subject of 'The Respiratory and Blood System'. Some of the teachers administered the instrument immediately after teaching the subject, while others used it later as preparation material for the end-of-year Matriculation exam in Biology.

The items were presented to students in a randomized order, to eliminate any dependency that might appear between questions, and to ensure that differences in student responses are not due to a specific ordering. Student explanations were typically 6-8 sentences long. Answers were written in Hebrew. Student responses were graded according to the grading rubrics by two experts. For each category, each response was graded as '1' if the property that is represented by this category was met, or '0' otherwise. First, a subset of the explanations (approximately 5% of the data) was graded together by both experts, and the application of the rubric was discussed (we note that this step yielded small modifications to the rubric). Second, another 20% of the data were graded independently by the two experts, and the level of agreement between them was evaluated. Inter-rater reliability was measured using Cohen's Kappa per category of the rubric and ranged from 0.86 to 0.98 (n=130). Third, the rest of the responses were graded by one of the experts. A representative example for each item, and its grading, is presented in Table 2 (the responses in their original form in Hebrew are included in the Appendix).

For each category, we paired optional feedback in case the category was graded as '0'. For example, in the first response presented in Table 2 (Anemia item), the student's response lacks reference to cellular respiration, and there is no connection between oxygen and ATP, and between ATP and exercise. Thus, feedback that may help the student to improve this answer would be: "You miss a connection between oxygen levels and the production of ATP. Can you explain how energy is produced in the cells?" and "What is the connection between ATP levels and the ability to exercise?"

The fraction of responses that were graded as correct in each of the categories is presented in Table 1. As can be seen, the Height item that was mentioned as requiring a more complex chain of reasoning (and as a consequence, its rubric includes more

Table 2 Examples of student responses to items 1, 2, and 3, and their human grading (translated from Hebrew). Categories that are not relevant to the item are marked with '–'

Student response	A	B	C	D	E	F	G	H	I	J	K	L	M
Red blood cells contain Hemoglobin to which oxygen binds. Oxygen moves from place to place with the help of red blood cells, so if there are fewer of them [red blood cells], less oxygen is transferred and enters the [body] cells. Exercise requires oxygen in the cells in order to form ATP molecules. Therefore if there is a lack of oxygen, less ATP is formed in the cells, which results in a feeling of fatigue and difficulty in exercising. (Anemia item)	–	–	–	1	1	1	1	0	0	0	1	1	0
CO gas is known to bind to Hemoglobin with a stronger tendency than oxygen. When CO binds to hemoglobin, it takes the oxygen's place, thus much less oxygen is transported from place to place and enters cells. Lack of oxygen in the cells leads to less production of ATP molecules. Since energy is required for physical exercise, the result is that the person gets tired quickly and has difficulty to exercise. (Smoking item)	–	–	1	0	1	1	0	0	0	0	1	1	1
In high altitudes, the amount of oxygen is low, which means there is less oxygen that enters the cells, therefore less cellular respiration takes place, and less ATP is produced. The group that came a month before [the competition] managed to adapt to the conditions of lack of oxygen, and more red blood cells were formed in the athletes' bodies. This means that more oxygen was transferred to the cells, and cellular respiration increased. The group that arrived a few days before [the competition] did not have time to adapt and produce red blood cells. (Height item)	1	1	0	0	1	1	0	1	1	0	1	1	0

categories), was more difficult for the students than the two other items, leading to very low scores (≤ 0.13) on some of its categories.

NLP and Machine Learning

Parsing and Word Embedding

To parse student responses, we used the publicly available Hebrew morphological parser developed by the National Institute of Testing and Evaluation (Cohen & Ben-Simon, 2011). The parser tokenizes the input texts and performs a morphological and syntactical analysis (parts of speech - POS). We then constructed a vocabulary of frequently-used (appear more than five times in the training set) morphemes and their POS. Our vocabulary length ranged between 300 to 600 morphemes depending on the

underlining training set. Each input text was then encoded as a sequence of indexes of the corresponding morphemes in the vocabulary, while removing stop words and encoding rare words with a special index. Finally, to make all the sequences of the same length, we padded the ones that are shorter than the max length with zeros (in our study the longest response was 370 tokens long, and the average length was 63 tokens).

Next, we performed Word Embedding, which aims to quantify semantic similarities between different words based on their tendency to occur in similar contexts (Mikolov et al., 2013). This step was based on analyzing the entire set of Hebrew Wikipedia articles. We used Gensim's Word2Vec CBOW algorithm (Rehurek & Sojka, 2010) over the latest dump of the Hebrew Wikipedia for creating word embeddings (of size 100).

Convolutional Neural Network

To classify (the parsed representations of) student responses we used a Convolutional Neural Network (CNN). The CNN consists of several layers. The overall architecture of the network is depicted in Fig. 2. Below we describe its layers in more detail.

Embeddings Layer

We initialized the weights of the embedding layer with the pre-trained word embeddings. The weights were later fine-tuned during model training. To prevent model overfitting we used a spatial dropout layer (Tompson et al., 2015) that randomly switches off the number of adjacent indexes in the embedding representation instead of using randomly chosen single indexes, as in the case of simple dropout. The latter is less effective for regularization (especially with a relatively small dataset as ours) as the remaining adjacent words (that are usually correlated semantically to the

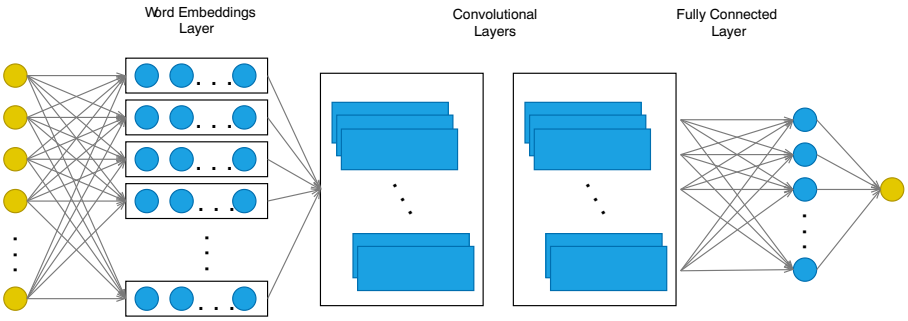

Fig. 2 The CNN consists of an Embedding Layer with embeddings of size 100 followed by a 1D Spatial Dropout Layer with a rate 0.2. Next we used 2 Convolutional Layers, each one with 128 1D filters and kernel size of 5. After the first convolution, we applied 1D Max-pooling Layer of size 2 and after the second convolution a 1D Global Max-pooling Layer. Finally, we used a Fully Connected Layer with 128 units and a dropout rate 0.5.

dropped word) continue to propagate to the subsequent layers values that are strongly correlated, decreasing the effect of the dropout.

Convolution and Pooling Layers

CNNs are widely used in various NLP tasks due to their ability to capture features based on certain ordered sequences of words no matter where they appear in the input text (Goldberg, 2016; Jacovi et al., 2018). In NLP, convolution and pooling architecture is constructed by applying a k-word convolutional filter (that is a non-linear learned function) to a word window of size k sliding over the input text and then pooling the max or average value to capture the existence of the feature in the text. Usually, a set of such filters followed by the max pooling operation is applied, resulting in d-size (d is the number of the filters in the set) dense representation of the entire text (Kim, 2014; Kalchbrenner et al., 2014; Johnson & Zhang, 2014). To optimize the number of convolutional filters (d) and their size (k), which are hyperparameters that need tuning, we used randomized grid search with $d = 32, 64, 128, 256$ and $k = 3, 5, 7$ (we also tested a combination of different filter sizes in one model (Zhang & Wallace, 2015)). The chosen architecture contains 2 layers of 128 convolutional filters of size 5.

Classification Layer

We modeled each rubric category as a binary classification task, using a fully connected dense layer with 128 units and a dropout rate of 0.5 for preventing model over-fitting (Srivastava et al., 2014). Finally, the output layer (with a Sigmoid activation) generates a binary grade.

Implementation

The CNN architecture was implemented using Keras API (Chollet & et al. 2015) with TensorFlow (Abadi et al., 2015) backend. We used the Adam optimizer (Kingma & Ba, 2014) with a learning rate of 0.001 to minimize the Binary Cross-Entropy loss. The models were trained with batch size 100 for 100 epochs and the model that achieved the best internal validation score was picked as the winning model.

Alternative NLP and Machine-Learning Approaches

In the preliminary stages of this research, we examined various machine learning architectures/approaches. Specifically, we tested LSTM (Hochreiter & Schmidhuber, 1997) and a combination of CNN and LSTM layers (Taghipour & Ng, 2016). We also tested classical machine learning approaches that combine Bag of Words (or Bag of Bigrams of Words) with SVM and Logistic regression classifiers. The CNN architecture with Gensim's Word2Vec word embeddings gave the best performance and was thus chosen for the experiments.

Baseline: Composite Analytic vs. Holistic Item Scores

As a baseline for validating the rationale that underlies our approach, we compared our analytic approach, which builds a dichotomous model per rubric category, with a holistic approach that builds a single model for each item, on the task of generating an item score. Below we describe the setup and results of this benchmark experiment.

Ground Truth and Datasets The ground truth for this benchmark is based on a per-item holistic $0-5$ grading scheme developed by two domain experts, which was computed as a weighted sum of the rubric categories of the item. This scheme was used to transform each student response, graded on each category, into a *holistic* grade. This yields additional holistic labels for each student response (in addition to the per-category labels of each response).

Single Models for Holistic Item Scores The baseline single item-level models use the same architecture as the per-category models, except for the classification layer that predicts $0-5$ categories instead of binary ones. The models were trained and evaluated on the holistic labels, using 10 times 5-fold cross-validation.

Analytic-based Composite Item Scores Using the dichotomous analytic bins to predict holistic scores is straightforward. Binary classification models were trained on the per-category labels. Then, they were applied to the Test set, and a composite score was computed by summing the predicted per-category classifications on each response and comparing this composite score with the holistic grade. The models were trained and evaluated using 10 times 5-fold cross-validation.

Benchmarking Results Table 3 compares the performance of the single item-level baseline models ('Holistic') and the Composite scores of the analytic models ('Composite'), on three metrics: Weighted Kappa with equal weights (WK), Weighted Kappa with square weights (WSK), and F1. As can be seen, scoring items by summing the analytic-based models outperform single item-level models on all items and all performance metrics.

Table 3 Performance of holistic item-level models and composite analytic-based models for item scores

	WK		WSK		F1	
	Holistic	Composite	Holistic	Composite	Holistic	Composite
Anemia	0.568	0.716	0.751	0.855	0.496	0.611
Smoking	0.643	0.734	0.803	0.872	0.557	0.613
Height	0.637	0.701	0.741	0.819	0.695	0.703

Evaluation Metrics

For answering the RQs of this study, we compared the automated scoring of the analytic bins to the expert grading, which served as ground truth. As evaluation metrics, we used Accuracy, *F1* statistics, and Cohen's Kappa. We followed the recommendation of Greer and Mark (2016) and Liu et al. (2016) and used Cohen's Kappa as the primary measure. The interpretation of what can be considered as a sufficient Kappa value may depend on the purpose (Liu et al., 2014). Williamson et al. (2012) argued that to be used in high-stakes testing, automatic graders should achieve Kappa>0.70. In a context that is more similar to ours – automated scoring aimed for formative purposes, Liu et al. (2014) suggests measuring the strength of agreement according to the more detailed distinction proposed by Landis and Koch (1977). They used the following interpretation of the kappa values: poor (< 0.00), slight ($0.00 - 0.20$), fair ($0.21 - 0.40$), moderate ($0.41 - 0.60$), good ($0.61 - 0.80$), and very good ($0.81 - 1$), which we adopt. Of course, Kappa, as any metric, has its disadvantages. When agreement among human experts is low, the AI system may have a latitude of choices that still result in high Kappa values, while in situations that the AI is superior to humans, Kappa values may be low even if the system performs well (Greer & Mark, 2016).

Results

We organize the results according to the research questions. As the rubrics were designed to support automated guidance based on the scoring of the categories, but can also be used to produce a holistic grade, we analyzed the performance of the scoring models both on the level of the categories and on the level of the entire item.

RQ1: Item-Level Models

We start with evaluating the performance of the item-level models, namely, models that are built and tested on the data of a single item. Overall, we have three items – Anemia, Smoking, and Height, with 10, 11, and 13 categories, respectively, yielding 34 trained models. The CNN models were evaluated using 5-fold cross-validation repeated 10 times. The results are presented in Table 4 (mean values; the SDs are reported in Table 7 in the Appendix). Overall, 62% of the classifiers achieved a good or very good level of agreement (light gray), 26% achieved a moderate level of agreement (gray), and 12% were fair or less (dark gray).

Analyzing per item, we can see that among the Anemia and Smoking items, 80% and 73% of the categories, respectively, reached a good or very good level of agreement. The classifiers trained on the categories of the Height item were much less successful, with only 38% reaching a good or very good level of agreement. Especially, the classifiers built for categories D, G, J, and M of item 3 achieved very low performance (dark gray, Kappa< 0.4). We believe that this is due to the data of these categories being highly imbalanced. As can be seen in Table 1, these categories were difficult for the students, with less than 13% of student responses marked as correct.

Table 4 Performance of the item-level models, colored according to the interpretation of their Kappa value: Good or very good (light gray), moderate (gray), and fair or less (dark gray)

	Anemia Item			Smoking Item			Height Item		
Category	Acc	F1	Kappa	Acc	F1	Kappa	Acc	F1	Kappa
A	-	-	-	-	-	-	.85	.80	.71
B	-	-	-	-	-	-	.88	.86	.76
C	-	-	-	.88	.59	.53	.87	.91	.70
D	.87	.83	.73	.85	.90	.60	.92	NA	.00
E	.85	.61	.52	.85	.71	.61	.81	.85	.60
F	.87	.86	.75	.90	.89	.80	.86	.91	.57
G	.87	.89	.72	.85	.89	.66	.89	.95	.32
H	.91	.92	.80	.88	.91	.71	.88	.96	.47
I	.97	.93	.93	.98	.91	.95	.97	.93	.93
J	.89	.96	.73	.84	.97	.46	.91	.98	.06
K	.90	.89	.80	.91	.91	.83	.86	.91	.60
L	.96	.93	.90	.96	.95	.91	.96	.97	.90
M	.85	.91	.57	.87	.91	.68	.92	NA	.00
Mean (D-M)	**.89**	**.87**	**.75**	**.89**	**.90**	**.72**	**.90**	**.93**	**.45**
K>.60 (D-M)			**80%**			**80%**			**20%**
Mean	**.89**	**.87**	**.75**	**.89**	**.87**	**.70**	**.89**	**.91**	**.51**
K>.60			**80%**			**73%**			**38%**

In the next subsection, we demonstrate that this issue may be addressed by combining data of several items.

RQ2: Between-Items Models

Next, we evaluated the ability of our models to generalize *between* items, namely, to score a certain category of a new item, after being trained for this category on data of a different item. The rationale is exploring scenarios where new items are graded using models built on data of existing ones, as a way to mitigate the 'cold start' problem. To evaluate this, we use the following three experiments. In the first two, we trained the models on the data of item 1 (item 2) and tested them on the data of item 2 (item 1). In the third, we trained the models on the data of item 1 and 2, and tested them on the data of item 3. In total, 30 models were built, for the 10 categories that are common among the three items (D-M). The models were evaluated by running each model 50 times with different random seeds. The performance is presented in Table 5 (mean values; the SDs are reported in Table 8 in the Appendix).

Comparing to the item-level models, we see the following (the comparison is made on the basis of categories D-M, which are common among all the items). As expected, using a model trained on one item to make inference on another item typically achieves lower performance. For the Anemia and Smoking items, the mean Kappa value decreased from 0.75 to 0.67, and from 0.72 to 0.70, respectively. The percentage of categories that reached at least a good level of agreement slightly decreased, from 80% to 70% for both. However, with respect to the Height item, in

Table 5 Performance of the between-items models, colored according to the interpretation of their Kappa value: Good or very good (light gray), moderate (gray), and fair or less (dark gray)

	Training Set Smoking Item Test Set Anemia Item			Training Set Anemia Item Test Set Smoking Item			Training Set Smoking and Anemia Item Test Set Height Item		
Category	Acc	F1	Kappa	Acc	F1	Kappa	Acc	F1	Kappa
D	.80	.78	.61	.79	.84	.55	.87	.92	.48
E	.83	.51	.43	.81	.60	.48	.68	.70	.38
F	.81	.77	.63	.89	.89	.79	.92	.95	.76
G	.83	.86	.65	.84	.88	.65	.93	.96	.71
H	.89	.91	.76	.87	.91	.68	.93	.96	.70
I	.98	.97	.95	.97	.97	.95	.98	.99	.95
J	.83	.89	.51	.86	.91	.62	.94	.97	.61
K	.89	.87	.78	.92	.92	.84	.93	.95	.80
L	.95	.93	.90	.97	.97	.94	.97	.98	.93
M	.83	.89	.52	.81	.88	.47	.95	.97	.61
Mean	**.86**	**.84**	**.67**	**.87**	**.88**	**.70**	**.91**	**.94**	**.69**
$K>.60$			**70%**			**70%**			**80%**

which the data on several of its categories were highly imbalanced, we see that using data of other items actually *improved* the results. Especially, in categories D, G, J, and M that were remarkably low, the Kappa increased significantly, to either moderate (D) or good (G, J, M). Only in category E, we see a decrease (from good to fair). Overall, the mean Kappa increased from 0.45 to 0.69, and the percentage of categories that achieved at least a good level of agreement increased dramatically from 20% to 80%.

RQ3: Instrument-Level Models

The last modeling that we evaluate is *instrument-level* models, namely, building a single model for each *category*, using the data of all the items that include this category – items 1-3 for categories D-M, items 2-3 for category C, and item 3 for categories A-B (the models for categories A-B are actually item-level models, thus are taken from Table 4, and are presented here for the completeness of the Table). Then, the quality of each model m that was built for category c was computed by measuring m's Kappa on each of the items that include category c. (The rationale for measuring each item separately was that in practice, grading and feedback are given on the item level, and items are required to achieve a satisfactory level of performance before they are deployed.) We evaluated all the models on a common held-out test set that consisted of 25% of the entire dataset. Per category, a model was built on the training set with a random seed and evaluated on the test data of each item separately. This

Table 6 Performance of the instrument-level (C-M) and item-level (A-B) models, colored according to the interpretation of their Kappa value: Good or very good (light gray), moderate (gray), and fair or less (dark gray)

	Anemia Item			Smoking Item			Height Item		
Category	Acc	F1	Kappa	Acc	F1	Kappa	Acc	F1	Kappa
A	-	-	-	-	-	-	.85	.80	.71
B	-	-	-	-	-	-	.88	.86	.76
C	-	-	-	.92	.75	.71	.87	.90	.71
D	.87	.82	.72	.92	.94	.80	.94	.97	.55
E	.85	.64	.55	.87	.75	.66	.84	.89	.64
F	.89	.88	.77	.92	.91	.83	.94	.97	.78
G	.88	.91	.74	.84	.88	.63	.94	.97	.71
H	.90	.92	.79	.93	.95	.84	.97	.98	.87
I	.99	.99	.99	.98	.98	.96	.99	.99	.97
J	.89	.93	.73	.89	.93	.66	.97	.98	.74
K	.93	.91	.85	.95	.95	.89	.97	.98	.92
L	.92	.87	.82	.97	.96	.94	.99	.99	.97
M	.85	.89	.65	.89	.92	.76	.94	.97	.53
Mean (D-M)	**.90**	**.88**	**.76**	**.92**	**.92**	**.80**	**.95**	**.97**	**.77**
K>.60 (D-M)			**90%**			**100%**			**80%**
Mean	**.90**	**.88**	**.76**	**.92**	**.90**	**.79**	**.93**	**.94**	**.76**
K>.60			**90%**			**100%**			**85%**

process was repeated 50 times. The results are presented in Table 6 (mean values; the SDs are reported in Table 9 in the Appendix).

As can be seen, instrument-level models (Table 6) achieved superior performance compared to the item-level models (Table 4): On the common categories (D-M), the mean kappa values of items 1, 2, and 3 increased from 0.75 to 0.76, from 0.72 to 0.80, and from 0.45 to 0.77, respectively. The amount of categories of each item that achieved at least a good level of agreement increased from 80% to 90%, from 80% to 100%, and from 20% to 80%, respectively. On the level of the entire item (all categories), the mean kappa values of items 1, 2, and 3 increased from 0.75 to 0.76, from 0.70 to 0.79, and from 0.51 to 0.76. The number of categories of items 1, 2, and 3 that achieved at least a good level of agreement increased from 80% to 90%, from 73% to 100%, and from 38% to 85%, respectively.

RQ4: Amount of Data that is Required

Finally, we seek to answer RQ4: How much labeled data is needed for our models to reach a good level of agreement with human grading? We evaluated this on the best performing modeling – instrument-level models for categories C-M, and item-level models for categories A-B (that exist in item 3 only).

Our experiment basically repeated the process for building Table 6, as described in the previous subsection, starting with a small subset of the training set and gradually increasing the subset till the entire training set is included. The held-out set was held constant for the entire process. We used a series of training sets of size $k =$

100, 200, ...1500. The choice of the k examples for the subset was done as follows. The entire training set was shuffled, and the examples were indexed from 1 to 1500. Then, for each k, the first k examples were taken as the training set. The rationale was to make sure that training sets of large k's include the training sets for smaller k's. This process was repeated 50 times, each time with a new shuffle. The results are presented in Fig. 3. The lower line (in blue) represents the mean Kappa values, and their SD, on the held-out set. The upper line (in red) represents the mean Kappa values, and their SD, on the training set. The dashed line marks the K=0.61 thresholds.

As can be seen, the models trained on each category tend to have a similar behavior across items. For example, category L is very easy to learn, and ∼200 training examples are enough to achieve almost perfect Kappa on this category for all the items. Category E is difficult to learn, and even after 1500 examples its Kappa is borderline on all the items. Only category D behaves differently across items – it is easy to learn on item 1 and item 2, but on item 3 even 1500 examples are not enough to achieve a good Kappa value. Overall, ∼500 examples are enough for ∼60% of the graders to achieve a good Kappa, and ∼900 examples are sufficient for ∼80% of the graders. After this point, adding more examples is not likely to have a substantial impact, and borderline models remain as such (e.g., Category J of item 2).

Summary of Results

Best Models We conclude that instrument-level models achieve the best performance. Measured as a level of agreement with human experts, they achieve Kappa values of 0.76, 0.79, and 0.76 on items 1, 2, and 3, respectively, with 90%, 100%, and 85% of the categories of these items scored with at least a good level of agreement. The mean Kappa value for all the items is 0.77, with 91% of the categories achieving at least a good level of agreement.

Amount of Data With respect to the amount of data that is needed, we conclude that ∼500 examples are enough to achieve a good level of agreement on more than half of the cases, ∼900 examples are sufficient for almost all of them, and after 900 examples we do not see substantial improvement in any of the cases.

Dicussion

This study aims to develop automated scoring models for open-ended items in Biology. We begin our discussion with analyzing the performance of *item-level* scoring models, namely, models that are trained and tested on the data of a single item (RQ1). This is the most narrow machine learning modeling. As demonstrated in the Results for RQ1, 62% of the models achieved a good level of agreement with the human grading. From the models that did not reach a good level of agreement, 12% demonstrated poor results ($\leq$ 0.4, meaning fair or less). The poorly performing models were all trained and tested on item 3 (Height), which is a complex item that was more difficult for the students. Specifically, these models were actually built on the most difficult categories ($\leq$ 13% correct, see Table 1), yielding imbalanced data sets.

Next, we turned our attention to modeling that generalizes *between* items – using scoring models trained on one item for making inference on responses to another item (RQ2; this was evaluated on categories D-M, which are common among all the items). The motivation to evaluate such models is 'practicality' – in interactive learning environments that are designated for learning (as opposed to testing environments for high-stakes exams), the content tends to be dynamic, so being able to score new items with models built for existing items can tremendously expedite the process of adding new items to the pool.

As expected, the performance of the between-items models on items 1 and 2 was inferior to the within-item models. Comparing Tables 4 and 5, we see a decrease from 0.75 to 0.67 and from 0.72 to 0.7 in the mean Kappa values of items 1 and 2,

Fig. 3 Mean and SD of the Kappa values as a function of the size of the Training set, computed on the Training (red) and Held-out sets (blue)

respectively, and a slight decrease in the number of categories on which the graders achieve at least a good agreement, from 80% to 70% in both items. However, on item 3, which its categories – especially the ones with the highly imbalanced data – were harder to learn, we actually see a significant improvement (see the Height item in Table 5). When using graders that were trained on data of item 1 and item 2, the mean Kappa value increased from 0.45 to 0.69, and the number of categories that achieved at least a good agreement increased from 20% to 80%.

But the best results were achieved by the instrument-level models (see Table 6). On the item level, the mean Kappa values were 0.76, 0.79, and 0.76 on items 1,2, and 3, respectively, and the percentage of categories that were graded with at least a good level of agreement are 90%, 100%, and 85%. Among the cells in Table 6, each representing a scoring task – applying a model to score a category within a certain item – our models achieved Kappa values in the range of 0.53-0.99. Overall, **91%** of the cells in Table 6 are scored with at least a good level of agreement.

Nevertheless, certain categories of the 'Height item' still had lower kappa values than others. This can be explained by the fact that these categories were underrepresented in students' answers, and also that they could be answered in more than one way, potentially making it more difficult for CNN to learn the correct patterns. We hypothesize that more examples that include the underrepresented categories, with the alternative possibilities for correct answers, could improve the kappa values of these categories. This hypothesis is strengthened by the fact that a significant improvement was achieved in the mean kappa values for these categories when including the data from the Anemia and Smoking items.

The fact that in our case instrument-level models achieved the best results contradicts the results of Nehm et al. (2012), which reported that item-level models achieved the best performance. This can be due to the fact that our grading approach decomposes the responses into several categories, most of them common to all the items in the instrument, thus making them more generalizable on the instrument level. Due to this generalizability, pooling data of several items helped the models to overcome the problem of imbalanced data in some of the categories (e.g., categories D, G, J, and M of item 3). Second, it probably helped the models to learn more sophisticated relations, by seeing more examples. Our results that demonstrate that analytic rubrics are especially beneficial for complex items, are inline with those of Jescovitch et al. (2021), who concluded that "It may be that analytic coding is beneficial to unpacking this additional complexity". However, our results underline an additional advantage of analytic rubrics, which is their superior generalizability, as models built on cross-item analytic bins, can be more easily applied across items and on the instrument level.

We note that the superior performance of instrument-level models is encouraging also because in practice it is typically easier to collect data on the instrument level (e.g., by administering it to several classes, and asking the teachers to grade the responses according to pre-defined rubrics). The fact that the pedagogic practice serves well the machine learning modeling, in this case, is a big plus.

Combining our results and previous work on the advantages of analytic rubrics, one can imagine a systematic approach for building large-scale automatic assessment systems that are based on accumulating models for *analytics bins*, rather than for

complete items. As more and more bins are trained, new questions that use various combinations of previously trained analytic bins can be scored without training new models.

Concerning how many human-scored responses are needed to achieve a good level of agreement (RQ4), our results yielded two key thresholds. First, ∼500 examples are enough in most cases. Second, going beyond ∼900 examples is not likely to produce a substantial impact on the Kappa value. These results have mainly two implications for practical implementation. First, as a rule of thumb, 500 examples (possibly pooling data of a few items) may be a reasonable starting point for training machine learning scoring models. Second, if a scoring model does not reach a good Kappa value after 900 examples, there is probably no point in adding more data, and either the grading rubric or its interpretation by the human graders should be revisited. Our findings strength the findings of Liu et al. (2016), who reported that typically 500–1,000 responses are needed to build reliable scoring models, and the results of Ha et al. (2011), who noted that models trained on a corpus of size n≈1000 did not necessarily perform better than models trained on a smaller corpus of size n≈500.

Our results compare well with the ones reported in a few highly relevant ASAS studies in science education. The mean Kappa that our graders achieved on items 1, 2, and 3, is 0.77. This places our results at the third quartile, just above the 0.72 median, in the range of Kappa values reported in the recent systematic literature review on applying machine learning in science assessment of Zhai et al. (2020). Similarly, compared to the results surveyed in Liu et al. (2016), our results fall into the upper half among the 6 studies reporting Kappa values and above the 2 studies reporting results with $F1$ score. In the context of results achieved in MRLs, Çınar et al. (2020) recently reported on the first application of NLP for automatic grading of open-ended questions, evaluated on Physics items in an undergraduate course. The mean $F1$ scores that they report (0.77 to 0.98) are slightly lower than ours – 0.88, 0.9, and 0.95 (mean $F1$ score for items 1,2, and 3).

Contextualizing our results in the more general domain of ASAS in science, we can compare our results to the state-of-the-art achieved by neural models on the large *SciEntsBank* dataset. On the 3-way SemEval-2013 Task (Dzikovska et al., 2013), the recent study of Sung et al. (2019) reported on mean *macro-average-F1* scores of 72.0 on "unseen answers (UA)", significantly lower than ours (and the other previously mentioned science education ASAS studies). The superior results of the science education studies may point to the importance of encoding nuanced and context-dependent information into the machine-learning modeling, but may also be explained by their relatively small scale, and the focus on certain types of questions. Either way, this emphasizes the need, and potential, of combining different streams of ASAS research.

We note that the majority of the studies surveyed in Zhai et al. (2020), and all the ones surveyed in Liu et al. (2016), have used off-the-shelf tools (e.g., c-rater (Leacock & Chodorow, 2003) or SIDE (Mayfield & Rosé, 2010)) that were optimized for this purpose. As our study is the first to explore this issue in Hebrew, we had to develop the whole NLP and machine learning pipeline, and there is certainly a lot of room for context- and task-specific improvements that will optimize our Hebrew NLP infrastructure for assessment in science education. Thus, we refer to our encouraging

results as a starting point and expect that future research and development will lead to significant improvements.

Furthermore, since Hebrew is an MRL, it poses considerable challenges for NLP. Indeed, except for the step within the morphological parser that analyses the tokens into morphemes (which are then transformed to POS and base form), our NLP pipeline does not use any MRL-specific modeling. However, as was noted by Tsarfaty et al. (2020), models that were developed with English in mind may not work well for MRLs, and parsing MRLs is typically more prone to errors. In Hebrew (and Arabic) there is also the issue of *vowelization*, which is not included in daily writing. This results in words with very different pronunciation and meaning having identical spelling, rendering additional ambiguity that the parser should resolve, which increases the complexity and error rate. At the beginning of this research, we could not anticipate the impact of these issues on the eventual result, and we are glad to be able to report for the first time in Hebrew that a relatively straightforward NLP and deep learning modeling can work well, given high-quality grading rubrics. That said, research on automatic assessment of scientific explanations in MRLs is scarce, and there is definitely a need for more research in order to understand how the special characteristics of MRLs impact different approaches for automatic grading. For example, it may be that MRLs would benefit more from structured grading rubrics, as they assist the machine learning models to capture the complexities of the language.

Limitations

The main limitation of the study is that its results are based on a small number of items from a single instrument on a single topic (cellular respiration) in Biology. While this amount of items is certainly not uncommon in studies on applications of language technologies to science education, further research is definitely required in order to evaluate the generalizability of the proposed approach to additional topics in Biology.

Another threat to the generalizability of our results relates to the size of the research sample, and whether it is representative. The classes that participated in the experiment were self-selected (by their teacher), and while the participating schools are spread in locations of varied socio-economic status all around the country, it is possible that the teachers who chose to participate are not a representative sample (for example, it is known that teachers of weak classes are less enthusiastic about joining educational experiments; however, to reduce this risk we notified the teacher in advance that the data will be collected anonymously).

Concerning the amount of data, our research used ∼2000 responses, which is a relatively small dataset for NLP tasks, even though quite typical to the science education domain (see for example in Nehm et al. (2012); Liu et al. (2014); Çınar et al. (2020)). Small datasets are especially limited in their ability to investigate (and address) questions of subgroup fairness based on race, gender, intellectual or physical disability, and more (Bridgeman et al., 2012; Madnani et al., 2017b). In this regard, however, it is important to mention that studies such as ours, which develop NLP-based solutions for under-represented languages may also help in reducing such biases against non-English speakers, who are in many cases required to take exams in their second language.

In addition, any practical application of automatic scoring methods to real-life settings should be prepared to deal with students developing sophisticated online cheating methods (Alexandron et al., 2017), or trying to game the system (Baker et al., 2008). For example, students may learn keywords or phrasal patterns that can improve their scores without learning the core material (Ding et al., 2020; Filighera et al., 2020), which may even result in biasing machine-learning algorithms in ways that affect their accuracy on the entire population (Alexandron et al., 2019).

Last, with respect to the 'ground truth', the grading rubrics, and their application to label the response data, are the result of a combined effort of Biology science education researchers and teachers. While this process synthesized science (and in particular, Biology) education assessment research, national and international biology assessment standards, and pedagogical expertise, its outcomes are obviously a reflection of the participants' collective wisdom, personal experience, and beliefs.

Implications to Science Education

As open-ended questions are necessary for capturing students' in-depth reasoning and argumentation, automated scoring systems that can provide students and teachers with timely and individualized guidance have significant implications for science education (Liu et al., 2016).

First, many existing NLP tools focus on writing mechanics rather than coherence or scientific accuracy (Tansomboon et al., 2017). Our study presents an approach for automatically evaluating student explanations on a more conceptual level, which also makes a fine-grained analysis of the micro-level components of the explanation, which takes into account the accuracy of the scientific explanation. It is done in Hebrew, but due to the complexity of this language, it is reasonable to assume that a similar approach can be applied to other languages as well. Therefore, our research can provide the scientific basis for developing computerized systems that can support science educators in teaching scientific writing, including specific guidelines for developing grading rubrics adapted for formative assessment of causal explanations.

Second, open-ended responses provide opportunities for assessing student ability to construct and communicate valid scientific explanations. Students benefit from guidance on their explanations and from the opportunity to revise them (Ryoo & Linn, 2014). Writing revisions provide students with the opportunity to reconsider ideas, a process that is beneficial to science learning (Tansomboon et al., 2017). Accordingly, our grading rubrics are specifically designed for providing *personalized* guidance based on students' performance, which we believe may assist them in revising and improving their written explanations. However, as no formative work was presented in this article, the efficacy of personalized guidance that is based on our feedback scheme should be studied and evaluated in future research, in authentic settings, by integrating the automated scoring algorithms into web-based tutors.

Finally, the ways in which our approach can reduce the burden involved in adding new items that require training data is another implication of this study. Our analysis suggests that some items can be generalized from previously learned items to new ones that have a similar structure and measure similar concepts. In this respect, future research should assess whether this generalizes to additional topics and types of open

ended questions in Biology, as well as in other scientific disciplines (e.g., physics and chemistry).

Summary

This paper presents the results of a study on automatic scoring of scientific explanations in Biology, conducted in an under-explored context – Morphologically-Rich Languages (MRLs), and specifically, Hebrew. In MRLs, each input token may consist of multiple lexical and functional units, making them particularly challenging for NLP. Our algorithms use Hebrew NLP and Deep Learning architecture to learn how to score student explanations according to analytic grading rubrics. The rubrics are designed to support automated guidance and are structured in a way that is amenable for machine learning. The results of our experiments show that our scoring algorithms can apply these rubrics to student responses, and achieve a high level of agreement with human experts, on par with previous work on automated assessment of scientific explanations in English, in which this area of research is well-established. This is the first study on automatic assessment of scientific explanations in Hebrew, and among the firsts to do so in MRLs. However, we believe that our approach for constructing grading rubrics for machine learning-based formative assessment of scientific explanations may be useful in other languages as well.

Appendix

M	L	K	J	I	H	G	F	E	D	C	B	A	תשובת התלמיד/ה
0	1	1	0	0	0	1	1	1	1	-	-	-	תאי הדם האדומים מכילים המוגלובין שאליו נקשר החמצן . החמצן עובר ממקום למקום בעזרת תאי הדם האדומים , לכן אם יש מעט מהם , פחות חמצן מועבר ונכנס לתאים . פעילות גופנית דורשת חמצן בתאים על מנת לייצר מולקולות ATP. לכן, כשיש פחות חמצן, פחות ATP נוצר בתאים , מה שגורם לתחושת עייפות וקושי לבצע פעילות גופנית (שאלת אנמיה).
1	1	1	0	0	0	0	1	1	0	1	-	-	CO הוא גז שידוע שנקשר להמוגלובין , בצורה יותר חזקה מאשר חמצן . כאשר CO נקשר להמולגובין , הוא תופס את מקומו של החמצן , ולכן פחות חמצן מועבר ממקום למקום ונכנס לתאים . מחסור בחמצן בתאים מוביל לירידה בהפקת מולקולות ATP . מאחר ואנרגיה דרושה לביצוע פעילות גופנית , התוצאה היאשהאדם נהיה עייף מהר , ויש לו קושי לבצע פעילות גופנית (שאלת עישון)
0	1	1	0	1	1	0	1	1	0	0	1	1	בגובה רב כמות החמצן היא נמוכה, מה שאומר שפחות חמצן נכנס אל התאים , ולכן מתרחשת פחות נשימה תאית , ונוצר פחות ATP . הקבוצה שהגיעה חודש לפני הצליחה להסתגל לתנאים של חוסר חמצן, ויותר תאי דם אדומים נוצרו בגופם של הספורטאים . זה אומר שיותר חמצן הועבר לתאים והנשימה התאית עלתה . הקבוצה שהגיעה מספר ימים לפני לא הספיקה להסתגל ולייצר תאי דם אדומים (שאלת שהייה בגובה)

Fig. 4 Examples of students' responses to the "Anemia", "Smoking", and "Height" items, in their original form in Hebrew (Hebrew is a right-to-left language)

Table 7 Performance of the item-level models, colored according to the interpretation of their Kappa value: Good or very good (light gray), moderate (gray), and fair or less (dark gray)

	Anemia Item			Smoking Item			Height Item		
Category	Acc	F1	Kappa	Acc	F1	Kappa	Acc	F1	Kappa
A	-	-	-	-	-	-	.85±.01	.80±.01	.71±.04
B	-	-	-	-	-	-	.88±.02	.86±.05	.76±.05
C	-	-	-	.88±.02	.59±.07	.53±.07	.87±.03	.91±.02	.70±.06
D	.87±.03	.83±.04	.73±.06	.85±.03	.90±.02	.60±.08	.92	NA	.00
E	.85±.03	.61±.10	.52±.11	.85±.02	.71±.07	.61±.07	.81±.03	.85±.03	.60±.08
F	.87±.03	.86±.04	.75±.07	.90±.02	.89±.03	.80±.05	.86±.04	.91±.02	.57±.18
G	.87±.03	.89±.03	.72±.06	.85±.03	.89±.02	.66±.07	.89±.02	.95±.01	.32±.18
H	.91±.03	.92±.02	.80±.07	.88±.03	.91±.02	.71±.07	.88±.03	.96±.02	.47±.24
I	.97±.01	.93±.01	.93±.03	.98±.01	.91±.01	.95±.02	.97±.01	.93±.01	.93±.03
J	.89±.03	.96±.02	.73±.07	.84±.04	.97±.02	.46±.16	.91±.01	.98±.01	.06±.19
K	.90±.02	.89±.03	.80±.05	.91±.02	.91±.02	.83±.04	.86±.03	.91±.02	.60±.09
L	.96±.02	.93±.02	.90±.04	.96±.01	.95±.01	.91±.03	.96±.02	.97±.01	.90±.04
M	.85±.04	.91±.02	.57±.14	.87±.03	.91±.02	.68±.09	.92	NA	.00
Mean (D-M)	**.89**	**.87**	**.75**	**.89**	**.90**	**.72**	**.90**	**.93**	**.45**
K>.60 (D-M)			**80%**			**80%**			**20%**
Mean	**.89**	**.87**	**.75**	**.89**	**.87**	**.70**	**.89**	**.91**	**.51**
sd	**.03**	**.03**	**.07**	**.02**	**.03**	**.07**	**.02**	**.02**	**.11**
K>.60			**80%**			**73%**			**38%**

Table 8 Performance of the between-items models, colored according to the interpretation of their Kappa value: Good or very good (light gray), moderate (gray), and fair or less (dark gray)

	Training Set Smoking Item Test Set Anemia Item			Training Set Anemia Item Test Set Smoking Item			Training Set Smoking and Anemia Item Test Set Height Item		
Category	Acc	F1	Kappa	Acc	F1	Kappa	Acc	F1	Kappa
D	.80±.02	.78±.02	.61±.04	.79±.04	.84±.04	.55±.06	.87±.03	.92±.02	.48±.07
E	.83±.02	.51±.09	.43±.09	.81±.02	.60±.05	.48±.05	.68±.05	.70±.07	.38±.07
F	.81±.03	.77±.06	.63±.07	.89±.01	.89±.01	.79±.02	.92±.01	.95±.01	.76±.05
G	.83±.01	.86±.02	.65±.03	.84±.01	.88±.01	.65±.02	.93±.01	.96±.01	.71±.04
H	.89±.01	.91±.01	.76±.03	.87±.02	.91±.01	.68±.05	.93±.01	.96±.01	.70±.08
I	.98±.01	.97±.01	.95±.01	.97±.01	.97±.01	.95±.02	.98±.01	.99±.01	.95±.01
J	.83±.04	.89±.02	.51±.15	.86±.01	.91±.01	.62±.03	.94±.01	.97±.01	.61±.10
K	.89±.01	.87±.01	.78±.02	.92±.01	.92±.01	.84±.02	.93±.01	.95±.01	.80±.03
L	.95±.01	.93±.01	.90±.01	.97±.01	.97±.01	.94±.01	.97±.01	.98±.01	.93±.02
M	.83±.01	.89±.01	.52±.06	.81±.04	.88±.02	.47±.18	.95±.01	.97±.01	.61±.04
Mean	**.86**	**.84**	**.67**	**.87**	**.88**	**.70**	**.91**	**.94**	**.69**
sd	**.02**	**.03**	**.05**	**.02**	**.02**	**.05**	**.02**	**.01**	**.05**
K>.60			**70%**			**70%**			**80%**

Table 9 Performance of the instrument-level (C-M) and item-level (A-B) models, colored according to the interpretation of their Kappa value: Good or very good (light gray), moderate (gray), and fair or less (dark gray)

	Anemia Item			Smoking Item			Height Item		
Category	Acc	F1	Kappa	Acc	F1	Kappa	Acc	F1	Kappa
A	-	-	-	-	-	-	.85±.01	.80±.01	.71±.04
B	-	-	-	-	-	-	.88±.02	.86±.05	.76±.05
C	-	-	-	.92±.01	.75±.05	.71±.06	.87±.01	.90±.01	.71±.02
D	.87±.02	.82±.02	.72±.03	.92±.02	.94±.01	.80±.05	.94±.01	.97±.01	.55±.09
E	.85±.01	.64±.04	.55±.05	.87±.01	.75±.02	.66±.03	.84±.02	.89±.02	.64±.04
F	.89±.02	.88±.02	.77±.04	.92±.01	.91±.01	.83±.02	.94±.01	.97±.01	.78±.03
G	.88±.01	.91±.01	.74±.03	.84±.01	.88±.01	.63±.03	.94±.01	.97±.01	.71±.04
H	.90±.01	.92±.01	.79±.03	.93±.01	.95±.01	.84±.03	.97±.01	.98±.01	.87±.06
I	.99±.01	.99±.01	.99±.01	.98±.01	.98±.01	.96±.01	.99±.01	.99±.01	.97±.01
J	.89±.02	.93±.01	.73±.04	.89±.01	.93±.01	.66±.05	.97±.01	.98±.01	.74±.04
K	.93±.01	.91±.01	.85±.02	.95±.01	.95±.01	.89±.02	.97±.01	.98±.01	.92±.03
L	.92±.01	.87±.01	.82±.01	.97±.01	.96±.01	.94±.01	.99±.01	.99±.01	.97±.01
M	.85±.02	.89±.01	.65±.05	.89±.02	.92±.02	.76±.05	.94±.01	.97±.01	.53±.17
Mean (D-M)	**.90**	**.88**	**.76**	**.92**	**.92**	**.80**	**.95**	**.97**	**.77**
K>.60 (D-M)			90%			100%			80%
Mean	**.90**	**.88**	**.76**	**.92**	**.90**	**.79**	**.93**	**.94**	**.76**
sd	**.01**	**.02**	**.03**	**.01**	**.02**	**.03**	**.01**	**.01**	**.05**
K>.60			90%			100%			85%

Acknowledgements The authors thank Cipy Hofman for her contribution. The research of GA and MA was supported by the Willner Family Leadership Institute for the Weizmann Institute of Science and the Iancovici-Fallmann Memorial Fund, established by Ruth and Henry Yancovich. TN is grateful to the Azrieli Foundation for the award of an Azrieli Fellowship.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., …Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Alexandron, G., Ruipérez-Valiente, J. A., Chen, Z., Muñoz-Merino, P. J., & Pritchard, D.E. (2017). Copying@ scale: Using harvesting accounts for collecting correct answers in a mooc. *Computers & Education*, *108*, 96–114.

Alexandron, G., Wiltrout, M. E., Berg, A., & Ruipérez-Valiente, J.A. (2020). Assessment that matters: Balancing reliability and learner-centered pedagogy in mooc assessment. In *Proceedings of the tenth international conference on learning analytics & knowledge* (pp. 512–517).

Alexandron, G., Yoo, L. Y., Ruipérez-Valiente, J. A., Lee, S., & Pritchard, D.E. (2019). Are mooc learning analytics results trustworthy? with fake learners, they might not be!. *International Journal of Artificial Intelligence in Education*, *29*(4), 484–506.

Allen, L. K., Jacovina, M. E., & McNamara, D.S. (2016). Computer-based writing instruction. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.) *Handbook of writing research, chapter 21*. 2nd edn. (pp. 316–329). Guilford Press.

Ariely, M., Nazaretsky, T., & Alexandron, G. (2020). First steps towards nlp-based formative feedback to improve scientific writing in hebrew. In A. N. Rafferty, J. Whitehill, C. Romero, & V. Cavalli-Sforza (Eds.) *Proceedings of the 13th international conference on educational data mining (EDM 2020)* (pp. 565–568).

Baker, R., Walonoski, J., Heffernan, N., Roll, I., Corbett, A., & Koedinger, K. (2008). Why students engage in "gaming the system" behavior in interactive learning environments. *Journal of Interactive Learning Research*, *19*(2), 185–224.

Berland, L. K., & Reiser, B. J. (2009). Making sense of argumentation and explanation. *Science Education*, *93*(1), 26–55.

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *The Phi Delta Kappan*, *80*(2), 139–148.

Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education*, *25*(1), 27–40.

Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, *25*(1), 60–117.

Chollet, F., et al. (2015). Keras. https://github.com/fchollet/keras.

Çınar, A., Ince, E., Gezer, M., & Yılmaz, Ö. (2020). Machine learning algorithm for grading open-ended physics questions in turkish. *Education and Information Technologies*, 1–24.

Cohen, Y., & Ben-Simon, A. (2011). The hebrew language project: Automated essay scoring & readability analysis. In *IAEA annual conference, Vienna, Austria*.

Cohen, Y., Levi, E., & Ben-Simon, A. (2018). Validating human and automated scoring of essays against "true" scores. *Applied Measurement in Education*, *31*(3), 241–250.

Ding, Y., Riordan, B., Horbach, A., Cahill, A., & Zesch, T. (2020). Don't take "nswvtnvakgxpm" for an answer–the surprising vulnerability of automatic content scoring systems to adversarial input. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 882–892).

Dzikovska, M. O., Nielsen, R. D., Brew, C., Leacock, C., Giampiccolo, D., Bentivogli, L., Clark, P., Dagan, I., & Dang, H.T. (2013). Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. Technical report, NORTH TEXAS STATE UNIV DENTON.

Filighera, A., Steuer, T., & Rensing, C. (2020). Fooling automatic short answer grading systems. In *International Conference on Artificial Intelligence in Education* (pp. 177–190). Springer.

Flor, M., & Cahill, A. (2020). *Automated scoring of open-ended written responses – possibilities and challenges*. Berlin: Springer Science.

Gerard, L. F., & Linn, M. C. (2016). Using automated scores of student essays to support teacher guidance in classroom inquiry. *Journal of Science Teacher Education*, *27*(1), 111–129.

Goldberg, Y. (2016). A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, *57*, 345–420.

Gomaa, W. H., & Fahmy, A. A. (2014). Automatic scoring for answers to arabic test questions. *Computer Speech & Language*, *28*(4), 833–857.

Graesser, A. C., McNamara, D. S., Cai, Z., Conley, M., Li, H., & Pennebaker, J. (2014). Coh-metrix measures text characteristics at multiple levels of language and discourse. *The Elementary School Journal*, *115*(2), 210–229.

Greer, J., & Mark, M. (2016). Evaluation methods for intelligent tutoring systems revisited. *International Journal of Artificial Intelligence in Education*, *26*(1), 387–392.

Ha, M., Nehm, R. H., Urban-Lurain, M., & Merrill, J.E. (2011). Applying computerized-scoring models of written biological explanations across courses and colleges: prospects and limitations. *CBE—Life Sciences Education*, *10*(4), 379–393.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112.

Heilman, M., & Madnani, N. (2015). The impact of training data on automated short answer scoring performance. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 81–85).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Horbach, A., Palmer, A., & Pinkal, M. (2013). Using the text to evaluate short answers for reading comprehension exercises. In *Second joint conference on lexical and computational semantics (* SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity* (pp. 286-295).

Israeli Ministry of Education, a. (2011). Syllabus of Biological Studies (10th-12th grade). State of Israel Ministry of Education Curriculum Center, Jerusalem, Israel.

Jacobs, K., Itai, A., & Wintner, S. (2020). Acronyms: identification, expansion and disambiguation. *Annals of Mathematics and Artificial Intelligence*, *88*(5), 517–532.

Jacovi, A., Sar Shalom, O., & Goldberg, Y. (2018). Understanding convolutional neural networks for text classification. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: analyzing*

and interpreting neural networks for NLP (pp. 56–65). Brussels: Association for Computational Linguistics.

Jescovitch, L. N., Doherty, J. H., Scott, E. E., Cerchiara, J. A., Wenderoth, M. P., Urban-Lurain, M., Merrill, J., & Haudek, K.C (2019a). Challenges in developing computerized scoring models for principle-based reasoning in a physiology context.

Jescovitch, L. N., Scott, E. E., Cerchiara, J. A., Doherty, J. H., Wenderoth, M. P., Merrill, J. E., Urban-Lurain, M., & Haudek, K.C. (2019b). Deconstruction of holistic rubrics into analytic rubrics for large-scale assessments of students' reasoning of complex science concepts. *Practical Assessment, Research, and Evaluation*, *24*(1), 7.

Jescovitch, L. N., Scott, E. E., Cerchiara, J. A., Merrill, J., Urban-Lurain, M., Doherty, J. H., & Haudek, K.C. (2021). Comparison of machine learning performance using analytic and holistic coding approaches across constructed response assessments aligned to a science learning progression. *Journal of Science Education and Technology*, *30*(2), 150–167.

Johnson, R., & Zhang, T. (2014). Effective use of word order for text categorization with convolutional neural networks. arXiv:1412.1058.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: reliability, validity and educational consequences. *Educational Research Review*, *2*(2), 130–144.

Kalchbrenner, N., Grefenstette, E., & Blunsom, P. (2014). A convolutional neural network for modelling sentences. arXiv:1404.2188.

Kampourakis, K., & Neibert, K. (2018). Explanation in biology education. In K. Kampourakis, & M. J. Reiss (Eds.) *Teaching biology in schools: Global research, issues and trends, chapter 19* (pp. 236–248). New York and Abingdon: Routledge.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1746–1751). Doha: Association for Computational Linguistics.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv:1412.6980.

Klebanov, B. B., Burstein, J., Harackiewicz, J. M., Priniski, S. J., & Mulholland, M. (2017). Reflective writing about the utility value of science as a tool for increasing stem motivation and retention–can ai help scale up? *International Journal of Artificial Intelligence in Education*, *27*(4), 791–818.

Klebanov, B. B., & Madnani, N. (2020). Automated evaluation of writing–50 years and counting. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 7796–7810).

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. biometrics. pp. 159–174.

Leacock, C., & Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, *37*(4), 389–405.

Li, H., Gobert, J., & Dickler, R. (2017), Automated assessment for scientific explanations in on-line science inquiry. International Educational Data Mining Society.

Litman, D. J. (2016). Natural language processing for enhancing teaching and learning. In *AAAI* (pp. 4170–4176).

Liu, O. L., Brew, C., Blackmore, J., Gerard, L., Madhok, J., & Linn, M.C. (2014). Automated scoring of constructed-response science items: Prospects and obstacles. *Educational Measurement: Issues and Practice*, *33*(2), 19–28.

Liu, O. L., Rios, J. A., Heilman, M., Gerard, L., & Linn, M.C. (2016). Validation of automated scoring of science assessments. *Journal of Research in Science Teaching*, *53*(2), 215–233.

Madnani, N., Loukina, A., & Cahill, A. (2017a). A large scale quantitative exploration of modeling strategies for content scoring. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 457–467).

Madnani, N., Loukina, A., Von Davier, A., Burstein, J., & Cahill, A. (2017b). Building better open-source tools to support fairness in automated scoring. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing* (pp. 41–52).

Maestrales, S., Zhai, X., Touitou, I., Baker, Q., Schneider, B., & Krajcik, J. (2021). Using machine learning to score multi-dimensional assessments of chemistry and physics. *Journal of Science Education and Technology*, *30*(2), 239–254.

Matthews, K., Janicki, T., He, L., & Patterson, L. (2012). Implementation of an automated grading system with an adaptive learning component to affect student feedback and response time. *Journal of Information Systems Education*, *23*(1), 71–84.

Mayfield, E., & Rosé, C. (2010). An interactive tool for supporting error analysis for text mining. In *Proceedings of the NAACL HLT 2010 Demonstration Session* (pp. 25–28).

Mayfield, E., & Rosé, C. P. (2013). Lightside: Open source machine learning for text. In *Handbook of automated essay evaluation: Current applications and new directions* (pp. 146–157). Routledge.

McNamara, D. S., Crossley, S. A., & Roscoe, R. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods*, *45*(2), 499–515.

McNeill, K. L., & Krajcik, J. (2008). Scientific explanations: Characterizing and evaluating the effects of teachers' instructional practices on student learning. *Journal of Research in Science Teaching*, *45*(1), 53–78.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv:1301.3781.

Moharreri, K., Ha, M., & Nehm, R.H. (2014). Evograder: an online formative assessment tool for automatically evaluating written evolutionary explanations. *Evolution: Education and Outreach*, *7*(1), 1–14.

National Research Council (NRC) (2012). A framework for K-12 science education: Practices, crosscutting concepts, and core ideas. Cambridge: The National Academies Press.

Nehm, R. H., Ha, M., & Mayfield, E. (2012). Transforming biology assessment with machine learning: Automated scoring of written evolutionary explanations. *Journal of Science Education and Technology*, *21*(1), 183–196.

Nehm, R. H., & Haertig, H. (2012). Human vs. computer diagnosis of students' natural selection knowledge: testing the efficacy of text analytic software. *Journal of Science Education and Technology*, *21*(1), 56–73.

Osborne, J. F., & Patterson, A. (2011). Scientific argument and explanation: a necessary distinction? *Science Education*, *95*(4), 627–638.

Padó, U. (2016). Get semantic with me! the usefulness of different feature types for short-answer grading. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical Papers* (pp. 2186–2195).

Pado, U., & Kiefer, C. (2015). Short answer grading: When sorting helps and when it doesn't. In *Proceedings of the fourth workshop on NLP for computer-assisted language learning* (pp. 42–50).

Rahimi, Z., Litman, D., Correnti, R., Wang, E., & Matsumura, L.C. (2017). Assessing students' use of evidence and organization in response-to-text writing: Using natural language processing for rubric-based automated scoring. *International Journal of Artificial Intelligence in Education*, *27*(4), 694–728.

Rehurek, R., & Sojka, P. (2010). Software Framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. Citeseer*.

Roschelle, J., Dimitriadis, Y., & Hoppe, U. (2013). Classroom orchestration: synthesis. *Computers & Education*, *69*, 523–526.

Roscoe, R. D., Varner, L. K., Crossley, S. A., & McNamara, D.S. (2013). Developing pedagogically-guided algorithms for intelligent writing feedback. *International Journal of Learning Technology 25*, *8*(4), 362–381.

Ross, L. N. (2020). Causal concepts in biology: How pathways differ from mechanisms and why it matters. The British Journal for the Philosophy of Science.

Ryoo, K., & Linn, M. C. (2014). Designing guidance for interpreting dynamic visualizations: Generating versus reading explanations. *Journal of Research in Science Teaching*, *51*(2), 147–174.

Sakaguchi, K., Heilman, M., & Madnani, N. (2015). Effective feature integration for automated short answer scoring. In *Proceedings of the 2015 conference of the North American Chapter of the association for computational linguistics: Human language technologies* (pp. 1049–1054).

Seddah, D., Tsarfaty, R., Kübler, S., Candito, M., Choi, J., Farkas, R., Foster, J., Goenaga, I., Gojenola, K., Goldberg, Y., & et al. (2013). Overview of the spmrl 2013 shared task: cross-framework evaluation of parsing morphologically rich languages. Association for Computational Linguistics.

Segal, A., Hindi, S., Prusak, N., Swidan, O., Livni, A., Palatnic, A., Schwarz, B., & et al. (2017). Keeping the teacher in the loop: Technologies for monitoring group learning in real-time. In *International Conference on Artificial Intelligence in Education* (pp. 64–76). Springer.

Sheinfux, L. H., Greshler, T. A., Melnik, N., & Wintner, S. (2015). Hebrew Verbal multi-word expressions. In *Proceedings of the 22nd International Conference on Head-Driven Phrase Structure Grammar, Nanyang Technological University, NTU, Singapore* (pp. 122–135).

Songer, N. B., & Gotwals, A. W. (2012). Guiding explanation construction by children at the entry points of learning progressions. *Journal of Research in Science Teaching*, *49*(2), 141–165.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929–1958.

Strobl, C., Ailhaud, E., Benetos, K., Devitt, A., Kruse, O., Proske, A., & Rapp, C. (2019). Digital support for academic writing: a review of technologies and pedagogies. *Computers & Education*, *131*, 33–48.

Sung, C., Dhamecha, T. I., & Mukhi, N. (2019). Improving short answer grading using transformer-based pre-training. In *International Conference on Artificial Intelligence in Education* (pp. 469–481). Springer.

Taghipour, K., & Ng, H. T. (2016). A neural approach to automated essay scoring. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 1882–1891).

Tang, K.-S. (2016). Constructing scientific explanations through premise–reasoning–outcome (PRO): an exploratory study to scaffold students in structuring written explanations. *International Journal of Science Education*, *38*(9), 1415–1440.

Tansomboon, C., Gerard, L. F., Vitale, J. M., & Linn, M.C. (2017). Designing automated guidance to promote productive revision of science explanations. *International Journal of Artificial Intelligence in Education*, *27*(4), 729–757.

Taras, M. (2005). Assessment – summative and formative – some theoretical reflections. *British Journal of Educational Studies*, *53*(4), 466–478.

Tompson, J., Goroshin, R., Jain, A., LeCun, Y., & Bregler, C. (2015). Efficient object localization using convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 648–656).

Tsarfaty, R., Bareket, D., Klein, S., & Seker, A. (2020). From spmrl to nmrl: What did we learn (and unlearn) in a decade of parsing morphologically-rich languages (mrls)? arXiv:2005.01330.

Tsarfaty, R., Sadde, S., Klein, S., & Seker, A. (2019). What's Wrong with hebrew nlp? and how to make it right. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP); system demonstrations* (pp. 259–264).

Tsarfaty, R., Seddah, D., Kübler, S., & Nivre, J. (2013). Parsing morphologically rich languages: Introduction to the special issue. *Computational Linguistics*, *39*(1), 15–22.

Wang, C., Liu, X., Wang, L., Sun, Y., & Zhang, H. (2021). Automated scoring of chinese grades 7–9 students' competence in interpreting and arguing from evidence. *Journal of Science Education and Technology*, *30*(2), 269–282.

Weston, M., Parker, J., & Urban-Lurain, M. (2013). Comparing formative feedback reports: Human and automated text analysis of constructed response questions in biology. In *NARST annual conference, Rio Grande, Puerto Rico*.

Williamson, D. M., Xi, X., & Breyer, F.J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, *31*(1), 2–13.

Wilson, J., Roscoe, R., & Ahmed, Y. (2017). Automated formative writing assessment using a levels of language framework. *Assessing Writing*, *34*, 16–36.

Woods, B., Adamson, D., Miel, S., & Mayfield, E. (2017). Formative essay feedback using predictive scoring models. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 2071–2080).

Yao, L., Cahill, A., & McCaffrey, D.F. (2020). The impact of training data quality on automated content scoring performance.

Yune, S. J., Lee, S. Y., Im, S. J., Kam, B. S., & Baek, S.Y. (2018). Holistic rubric vs. analytic rubric for measuring clinical performance levels in medical students. *BMC Medical Education*, *18*(1), 1–6.

Zesch, T., Heilman, M., & Cahill, A. (2015). Reducing annotation efforts in supervised short answer scoring. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 124–132).

Zhai, X. (2021). Practices and theories: How can machine learning assist in innovative assessment practices in science education. *Journal of Science Education and Technology*, *30*(2), 139–149.

Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., & Shi, L. (2020). Applying machine learning in science assessment: a systematic review. *Studies in Science Education*, *56*(1), 111–151.

Zhang, Y., & Wallace, B. (2015). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. arXiv:1510.03820.

Zhu, M., Lee, H.-S., Wang, T., Liu, O. L., Belur, V., & Pallant, A. (2017). Investigating the impact of automated feedback on students' scientific argumentation. *International Journal of Science Education*, *39*(12), 1648–1668.

Zhu, M., Liu, O. L., & Lee, H. -S. (2020). The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers & Education*, *143*, 103668.

Zhu, M., Liu, O. L., Mao, L., & Pallant, A. (2016). Use of automated scoring and feedback in online interactive earth science tasks. In *2016 IEEE Integrated STEM Education Conference (ISEC)* (pp. 224–230). IEEE.

Affiliations

Moriah Ariely[1] · Tanya Nazaretsky[1] · Giora Alexandron[1]

Tanya Nazaretsky
tanya.nazaretsky@weizmann.ac.il

Giora Alexandron
giora.alexandron@weizmann.ac.il

[1] Weizmann Institute of Science, Rehovot, Israel