Personalized Automated Formative Feedback Can Support Students in Generating Causal Explanations in Biology

Moriah Ariely, Tanya Nazaretsky and Giora Alexandron
moriah.ariely@weizmann.ac.il, tanya.nazaretsky@weizmann.ac.il, giora.alexandron@weizmann.ac.il
Weizmann Institute of Science

Abstract: The ability to formulate scientific explanations is considered a major goal in science education. However, explaining scientific phenomena is a challenging task, and research has shown that many students are unable to construct proper explanations. And while personalized formative feedback can greatly assist students in closing the gaps between their current and the expected level of reasoning, teachers rarely find the time that is required to provide it. We present a novel framework for designing and providing formative feedback on causal explanations in biology, which can be generated automatically using NLP-based algorithms. Results from a controlled experiment showed that applying this framework led to significant improvement in student explanations. Based on these findings, we believe that our framework provides an effective approach for aligning between automated NLP-based analysis and formative assessment of biological explanations in the science classroom.

Introduction
Enhancing students’ ability to formulate scientific explanations is considered a major goal of science education (National Research Council (NRC), 2012). However, research has shown that many students are unable to construct proper explanations (e.g., Nehm et al., 2012).

Constructed-response (CR) items are an important tool to elicit students’ in-depth understanding of scientific concepts (Nehm et al., 2012), and students were found to benefit from guidance on their explanations and the opportunity to revise them (Ryoo & Linn, 2014; Tansomboon et al., 2017). However, proper assessment of CR items is a costly, time-consuming process, which requires developing grading rubrics and applying them to student responses. Teachers are struggling to find the time to devote to this process, leading to considerable delays in the feedback students receive on their CRs, and affecting its quality (Ariely et al., 2022).

Technology holds much promise for improving this process. Research on automated assessment in science education has shown that machine-learning (ML) and Natural Language Processing (NLP) can be used to construct scoring algorithms and provide immediate feedback to teachers and students (Zhai et al., 2020). However, few studies have attempted to leverage the power of automated scoring techniques for providing feedback on CR items. Moreover, research on the educational impact of formative assessment with automated scoring technologies in real classrooms – where most of the learning occurs – is scarce (Zhu et al., 2020). Our research aims for this gap.

Causal explanations in biology
The majority of explanations in the science classroom are attempts to provide an account that specifies what happened and how it occurred, namely, constructing causal accounts of phenomena (Osborne & Patterson, 2011). Causal explanation should include the sequential stages from input to output of the underlying causal event leading to a phenomenon, and also how one or more factors behave to give rise to a phenomenon (Krist et al., 2019). Depending on the phenomenon to be explained and the context in which it occurs, the underlying causality can be a simple linear relation between contiguous events, or it can consist of complex relations, with several causal factors interacting with each other (Grotzer, 2003). Studies show that students have difficulties in presenting the causal story in which event results from interactions among multiple components (Grotzer, 2003).

Automated formative assessment
Automated formative assessment of CR 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 (e.g., Tansomboon et al., 2017). Application of ML and NLP are increasingly applied to automated content scoring of CR items (Zhai et al., 2020). Automated scoring techniques have many advantages for assessment such as objectivity, standardization and efficiency (e.g., Nehm et al. 2012; Tansomboon et al., 2017). Also, ML-based assessment was found to promote significant learning gains (e.g., Tansomboon et al., 2017; Zhu et al., 2020). This makes ML-based assessment a cost-effective way for assessing students’ explanations.
Research goals and questions
In this study, we present a novel approach and method for designing and providing formative feedback on causal explanations in biology. Our method “breaks down” the expected explanation and reveals the components of the causal explanation. These components are then used to build fine-grained analytic rubrics that are designed for formative assessment purposes and aimed at capturing conceptual understanding.

We recently showed that NLP-based ML scoring models, that were trained on student response data that were human-labeled according to our analytic grading rubrics, can reach a high level of agreement with human experts on the task of scoring scientific explanations (Ariely et al., 2022). In the present study, we applied the automated scoring models to provide personalized feedback to students based on their performance in an authentic classroom scenario. While performed on a case study, our approach is directly applicable to causal explanations in other subjects in biology, and may be relevant to other disciplines as well.

The research questions that we seek to answer are: (1) how do students’ responses to open-ended questions change after receiving personalized feedback and- (2) if our feedback on students’ response to one item affects their response to a different item that measures the same concepts in a different context?

Methods
Participants and data collection
The main data set included 10th-grade students from five classes in the same urban school at the center of Israel (N=115). Due to ethical regulations, answers were collected anonymously, and students were requested to fill only their grade, gender, and geographic location. When needed, the class teacher numbered her students so that she could provide each student their personalized feedback, and track their performance.

Procedure
The constructed response items
Two constructed response items were presented to the students as part of their regular biology studies, and deal with the effect of anemia and smoking on physical activity (hereafter referred to as Anemia and Smoking items). Both items are typical open-ended questions in biology and require a causal explanation to explain the phenomena.

Groups description
The data set (N=115) included two groups as follows: One group included three classes in which the students submitted their responses to the Anemia item, and received personalized feedback on their response ~24 hours after submission. Then, the students were asked to submit their responses to the same item again, and also to another item- the Smoking item. Seventy-three students answered the Anemia item, but 19 of them did not submit their answer to the Anemia and Smoking item after receiving feedback, and were, thus, excluded from the analysis. Therefore, the final sample for this group was N=54 (hereafter the “Treatment group”). A second group included two classes in which the students answered both items one after the other, in a random, unknown order, submitted the items together, and did not receive any feedback on their responses (N=61; hereafter the “Control group”).

Automated scoring models and analytic grading rubrics
The students’ responses to the Anemia and Smoking items were automatically graded using automated scoring models, developed for these items in prior research (Ariely et al., 2022). Analytic grading rubrics were developed by the first author (a domain expert) and validated by a second domain expert and five experienced biology teachers. The rubrics decompose a correct explanation into a collection of binary categories, each representing an essential property that the student explanation should include (see Table 1). A detailed description of the development and performance of the automated scoring models can be found in Ariely et al. (2022).

Personalized feedback on students’ responses
The feedback on students’ responses (in case they were not fully correct) included a short paragraph followed by a figure that visualizes what is missing from the student’s explanation and the student’s score out of 5 points. The content of the feedback was predefined according to the different types of reasoning that are revealed by the analytic grading rubric. The feedback was sent to the class teacher who sent it to her students, and also provided a short explanation about the feedback format.

After the students in the Treatment group submitted their responses (following receiving the feedback), they answered a short questionnaire about how they used the feedback, and what part of the feedback was most
useful to them. This questionnaire served two purposes: First, characterizing what the students perceived as being most useful for them. Second, it also provided evidence that students actually read the feedback handed to them.

Table 1.
The categories of the analytic grading rubrics for the Anemia and Smoking items. The (+) and (−) signs represent which category is relevant to each item.

<table>
<thead>
<tr>
<th>Category – present (+) not present (−)</th>
<th>Anemia item</th>
<th>Smoking item</th>
</tr>
</thead>
<tbody>
<tr>
<td>a CO-Oxygen competition</td>
<td>−</td>
<td>+</td>
</tr>
<tr>
<td>b Oxygen is transported to the cells by RBC</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>c Change in cell oxygen levels</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>d Oxygen is a reactant in cellular respiration</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>e Change in cellular respiration rates</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>f Use of term (cellular respiration)</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>g Energy is produced in cellular respiration</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>h Change in energy production rates</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>i Use of term (energy/ATP)</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>j Energy is consumed by muscles/cells</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

Data analysis

To evaluate the impact of the feedback, we analyzed students’ performance in a real classroom situation, by assessing their performance in two different, but conceptually similar, open-ended questions, before and after receiving feedback. This relies on the fact that our assessment and analytic grading rubrics allows us to compare items that measure the same concepts in different contexts.

The students’ responses were automatically scored 0 or 1 for each category in the analytic grading rubric, and then these scores were combined to form a composite, holistic score for each item (out of 100%). We analyzed students’ performance on the Anemia and Smoking items by comparing their holistic scores for each item. We used the Wilcoxon signed-rank test (for dependent samples), and Mann-Whitney U-test (for independent samples) to analyze differences between students’ average holistic scores (followed by a Bonferroni correction, as the Anemia item was compared twice for the Treatment and Control groups).

Results

We analyzed students’ performance on the Anemia and Smoking items by calculating their holistic score. We found that while both groups performed similarly on the pretest (Anemia items; two-tailed Mann-Whitney U-test, $z = 0.35305, p = 0.72634$, Table 3), the control group did not improve on the posttest (Anemia vs. Smoking item, two-tailed Wilcoxon signed-rank test, $z$-score $= -1.2039, p = 0.2301$, Table 3). However, the treatment group improved compared to its pretest results; the students’ performance in the Anemia item improved significantly after they received personalized feedback on the same item (Anemia item pre vs. post, two-tailed Wilcoxon signed-rank test, $z$-score $= -5.3527, p < .00001$; Table 2), and also on a different item (Anemia vs. Smoking item, two-tailed Wilcoxon signed-rank test, $z$-score $= -2.809, p = 0.005$, Table 3). Moreover, the treatment group (with feedback) performed better than the control group (no feedback) on the posttest (Smoking items; two-tailed Mann-Whitney U-test, $z$ = -2.1491, $p = 0.031$, Table 3).

Table 2.
The Treatment group’s average scores in the Anemia item before and after feedback

<table>
<thead>
<tr>
<th>Treatment group (N=54)</th>
<th>Anemia item (before feedback)</th>
<th>Anemia item (after feedback)</th>
<th>Wilcoxon signed-rank test (Anemia pre vs. Anemia post)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>51.9</td>
<td>80.0</td>
<td>$z$-score $= -5.3527, p &lt; .00001$</td>
</tr>
</tbody>
</table>

Table 3.
The Treatment and Control groups’ scores to the Anemia vs. Smoking items before and after feedback

<table>
<thead>
<tr>
<th></th>
<th>Anemia item</th>
<th>Smoking item</th>
<th>Wilcoxon signed-rank test (Anemia vs. Smoking)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment group (N=54)</td>
<td>51.9</td>
<td>66.3</td>
<td>$z$-score $= -2.809, p &lt; .01$</td>
</tr>
<tr>
<td>Control group (N=61)</td>
<td>54.1</td>
<td>57.2</td>
<td>Not significant</td>
</tr>
</tbody>
</table>

*Mann-Whitney U- Test (Treatment vs. Control)

Mann-Whitney U-Test (Treatment vs. Control) Not significant $z$-score $= -2.1491, p < .05$

*Bonferroni correction (significant if $p < .025$)
Discussion
In this study, we present a novel approach and method for designing and providing automated formative feedback on causal explanations in biology. Our analysis shows that after receiving feedback on their response to one item (the Anemia item), students improved their responses to the same item, but also to a different item that measures the same underlying biology concepts (the Smoking item).

Studies on formative feedback stress that in order to be effective, the feedback should be immediate, individualized, and offer concrete guidance for improvement (Tansomboon et al., 2017). As presented in this study, machine learning algorithms that automatically score scientific explanations can be used to identify gaps in students’ reasoning and provide them with timely and individualized feedback. In addition, our grading rubrics were specifically designed for providing personalized guidance based on the students’ performance, which we believe assisted in students’ revisions of their constructed responses.

More importantly, our approach for assessment can provide teachers and students with insights about how to generate and evaluate scientific explanations, because it makes the rationale of the scientific explanation and its components explicit, which can promote students’ learning of scientific explanations. It is claimed that crafting complex causal explanations for biological phenomena, promotes students’ understanding of the phenomenon and the underlying ideas (Grotzer & Bell Basca, 2003). Also, when students are involved in identifying causal explanations for a phenomenon they are prompt to infer about the underlying causes and to discover the structural connections of phenomena in the world (de Andrade et al., 2019).

Our approach can also provide researchers with a method for developing rubrics and models for automated formative assessment. Nevertheless, as the main limitation of this research is that it is based on a case study with two items on a certain biological phenomenon, further research is required to validate the generalizability of this approach to larger sets of items and of other topics within the biology curriculum.

References


Acknowledgments
The work of MA was partially supported by the Israeli Council for Higher Education (CHE) via the Weizmann Data Science Research Center. TN is grateful to the Azrieli Foundation for the award of an Azrieli Fellowship.