Comparing Human and Automated Evaluation of Open-Ended Student Responses to Questions of Evolution

Michael J. Wiser[1,2], Louise S. Mead[1,2,3], James J. Smith[1,2,3,4,5], and Robert T. Pennock[1,2,4,6,7]

[1]BEACON Center for the Study of Evolution in Action

[2]Program in Ecology, Evolutionary Biology and Behavior, Michigan State University, East Lansing, MI, USA

[3]Department of Integrative Biology, Michigan State University, East Lansing, MI, USA

[4]Lyman Briggs College, Michigan State University, E. Lansing, MI, USA

[5]Department of Entomology, Michigan State University, East Lansing, MI, USA

[6]Department of Philosophy, Michigan State University, East Lansing, MI, USA

[7]Department of Computer Science & Engineering, Michigan State University, East Lansing, MI, USA

mwiser@msu.edu

Abstract

Written responses can provide a wealth of data in understanding student reasoning on a topic. Yet they are time- and labor-intensive to score, requiring many instructors to forego them except as limited parts of summative assessments at the end of a unit or course. Recent developments in Machine Learning (ML) have produced computational methods of scoring written responses for the presence or absence of specific concepts. Here, we compare the scores from one particular ML program – EvoGrader – to human scoring of responses to structurally- and content-similar questions that are distinct from the ones the program was trained on. We find that there is substantial inter-rater reliability between the human and ML scoring. However, sufficient systematic differences remain between the human and ML scoring that we advise only using the ML scoring for formative, rather than summative, assessment of student reasoning.

Background

The central importance of evolution to teaching and learning in the biological sciences has been clearly established in all science education reform (States, 1900; Brewer and Smith, 2011). Adequate formative assessment instruments – administered during the course of instruction to gauge student understanding and reasoning in order to provide feedback for future instruction, instead of to assign a grade at the end of a unit – that measure student understanding of evolutionary concepts (Bishop and Anderson, 1990; Anderson et al., 2002), however, have until recently been rather limited (Nehm and Schonfeld, 2008). Part of the challenge in designing an effective instrument comes from the fact that student understanding of evolutionary concepts is complex, and constantly changing. Studies find that students hold both scientifically accurate and naive or non-scientific explanations simultaneously (Andrews et al., 2012; Hiatt et al., 2013) and that accurately identifying alternative conceptions can be difficult (Rector et al., 2012). Data also suggest students reason differently than experts, especially in response to different contextual elements of the sample questions. Undergraduates employ more naive concepts when applying explanations of natural selection to plants as compared to animals; trait loss as compared to trait gain; and unfamiliar taxa as compared to familiar taxa (Nehm and Ha, 2011). Furthermore, ascertaining the meaning of student responses is often very difficult. One study found that 81 percent of students incorporated lexically ambiguous language in their responses to open ended questions about evolutionary mechanisms (Rector et al., 2012).

Despite these challenges, assessing student knowledge is important, particularly in evaluating pedagogical practices designed to improve student understanding. In an effort to identify effective assessment strategies, we have been investigating the applicability of a new tool, EvoGrader (Moharreri et al., 2014).

Open-ended student responses can provide a wealth of data about student reasoning. Unfortunately, they can also be time- and labor-intensive to score. One study found that it took an average of four minutes for a human grader to score a single response for the nine ideas we analyze in this study (Moharreri et al., 2014). For even a class of 30 students, scoring five such questions would take ten hours, which quickly becomes prohibitive. If an instructor wants to get a general sense of student understanding on a formative assessment, a more rapid method is highly desirable.

An appealing potential solution to this problem would be if instructors had an automated system that was sufficiently sophisticated to evaluate student answers to such open-ended questions. Of course, this is not a simple task. Even setting aside the difficulty of parsing open-ended natural language responses in general, one still has the further problem of interpreting the appropriateness of answers in relation to content knowledge and overarching concepts. For instance, a science teacher may want to know whether a student's response demonstrates incorrect naive notions or whether it demonstrates concrete scientific understanding. Machine Learning systems have begun taking the first steps to accomplishing this difficult task.

Use of Machine Learning in Education

There is growing interest in using tools and techniques from Machine Learning in the classroom environment (Butler et al., 2014). In fact, a chapter has been written about using Machine Learning in educational science (Kidziński et al., 2016) within the context of a book on educational technologies. One area of particular interest is language processing. Machine learning techniques have been used to classify instructor questions according to Bloom's taxonomy (Yahya et al., 2013). Perhaps the biggest use of Machine Learning in an educational environment is in the automated scoring of student writing (reviewed in (Nehm et al., 2012b)).

One domain-specific example of ML techniques in language processing is provided by the web portal EvoGrader, discussed below. EvoGrader was designed to assess student understanding of natural selection, using a particular set of questions, consisting of a brief scenario and asking the students how a biologist would explain this scenario of evolutionary change or patterns. Our study seeks to measure how similar of scores this ML procedure provides to human scoring for questions on which the application has not been trained but which are written in the same style.

EvoGrader

EvoGrader (`http://www.evograder.org`) is a free, online service that analyzes open-ended responses to questions about evolution and natural selection, and provides users with formative assessments. It is described in detail in (Moharreri et al., 2014), but a brief description follows.

EvoGrader works by supervised machine learning. Participants (n=2,978) wrote responses to ACORNS assessment items (Nehm et al., 2012a) and ACORNS-like items (Bishop and Anderson, 1990), generating 10,270 student responses. These items consist of a prompt describing a short scenario relevant to natural selection, and ask students to write how a biologist would explain this situation. Participants spanned many different levels of expertise, including non-majors, undergraduate biology or anthropology majors, graduate students, postdocs, and faculty in evolutionary science. Each response was scored independently by two human raters for each of six Key Concepts (KC) and three Naive Ideas (NI) (see Box 1). These consensus scores were used to train EvoGrader, based on the supervised machine learning tools of LightSIDE (Mayfield and Rosé, 2013). LightSIDE provides feature extraction, model construction, and model validation, based on the human-scored responses.

EvoGrader's authors chose different methods to optimize the scoring algorithm for feature extraction for the 9 scoring models (one model for each concept) – all considered the dictionary of words used in a particular response, and reduced words to their stems; most removed high frequency low information words (e.g., the, of, and, it); some also included pairs of consecutive words (e.g., "had to", "passing on"), or removing misclassified data (see Moharreri et. al. (Moharreri et al., 2014) Table 2 for details).

After feature extraction, each response was converted to a set of vectors containing frequencies of words or word pairs. These vectors were then passed to a binary classifier, which underwent Sequential Minimal Optimization (SMO) (Platt 1999) for each of the 9 models. The SMO training algorithm iteratively assigned weights to words in the written responses until the model was able to match the human scores within a certain margin of error. The models were then validated with 10-fold cross-validation, using 90% of the data to generate a model and the remaining 10% of the data to validate it, and then repeating this procedure for a total of 10 times such that each 10% of the data was used for validation exactly once and model generation 9 times. The authors averaged these models to get the final models used by the program, assessing whether they met quality benchmarks (90% accuracy and kappa coefficients ≥ 0.8) defined by the creators, and adjusting the training until the models did.

EvoGrader uses these validated models to score new responses from web users. Users must upload data in a specific format, which the portal verifies. If the data is formatted correctly, EvoGrader then evaluates each response using the existing validated models, and provides both machine scored data in a downloadable .csv format and a variety of web visualizations of the data. (Fig. 1)

Methods

Student data

We administered pre-instruction and post-instruction tests consisting of two questions (see Box 1) about evolution to students in an Introductory Cell and Molecular Biology course in the fall semester of 2014. Both questions asked students about how evolutionary processes occur. Question 1 asks about an evolutionary gain of antibiotic resistance in a population, while question 2 asks about the evolutionary loss of toxicity in a mushroom population. Completed pre- and post-test responses were obtained from 34 students for question 1 and from 36 students for question 2.

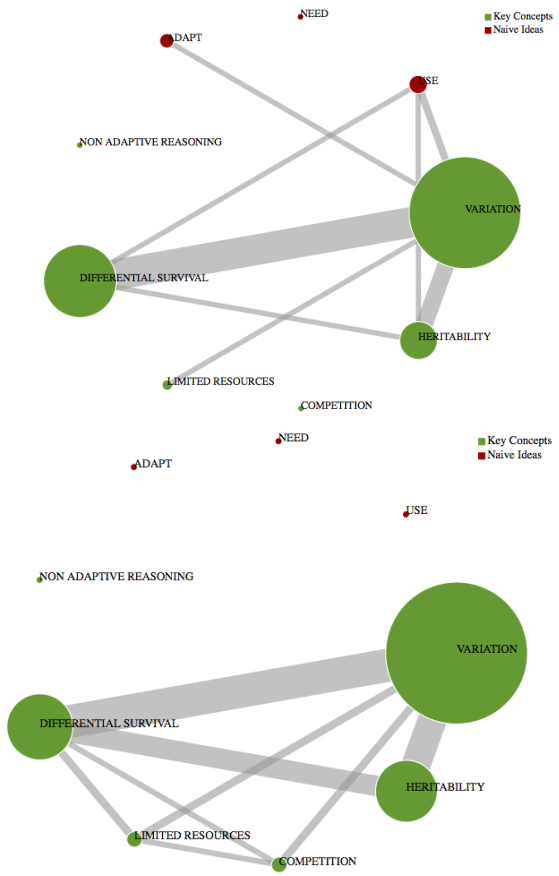

Figure 1: Concept maps produced by EvoGrader for the pre-instruction (upper panel) and post-instruction (lower panel) analysis of Question 1 (see Box 1). Sizes of the circles indicate percentage of responses scored as containing that concept; widths of the lines connecting concepts shows frequency of co-occurrence of those concepts.

Box 1

We evaluated student responses to two prompts:

Question 1: Explain how a microbial population evolves resistance to the effects of an antibiotic.

Question 2: A species of mushroom contains a chemical that is toxic to mammals. How would biologists explain the initial occurrence and increase in frequency of a number of individuals in the population that no longer produce this toxin?

We scored each response for whether it contained each of the following concepts:

Key Concepts:

- *Variation*: The presence and causes (mutation/recombination/sex) of differences among individuals in a population.
- *Heritability*: Traits that have a genetic basis and are able to be passed on from parent to offspring.
- *Competition*: A situation in which two or more individuals struggle to get resources which are not available to everyone.
- *Limited Resources*: Required resources for survival (food, mates, water, etc) which are not available in unlimited amounts.
- *Differential Survival*: Differential survival and/or reproduction of individuals.
- *Non-adaptive Ideas*: Genetic drift and related non-adaptive factors contributing to evolutionary change.

Naive Ideas:

- *Adapt*: Organisms/populations adjust or acclimate to their environment.
- *Need*: Organisms gain traits or advantage in response to a need or a goal to accomplish something.
- *Use/Disuse*: Traits are lost or gained due to use or disuse of traits.

Further, human evaluators determined whether or not a response answered the question asked; if the response did not, no credit was given for Key Concepts. For example, consider this student response:

> *Similar to above, some kind of mutation for the poison and those plants were not eaten so they were able to reproduce and pass thoses* [sic] *genes on to future generations. The population of poisonous mushrooms would soon outnumber non-poisonous ones since poisonous mushrooms are less likely to be eaten. Over time, animals would learn to stay away from teh* [sic] *mushroom simply be [sic] appearance, so the toxin would no longer be needed.*

Although this answer demonstrates adaptive reasoning about the origin of toxic mushrooms, the question was about the loss of toxin in this population, not the origin of the toxin. Only the last sentence addresses the loss of the toxin, and it does not demonstrate any of the Key Concepts.

Data

Data files containing all student responses, scoring, and data analysis may be found at `https://github.com/mjwiser/ALife2016`

Scoring responses

We used EvoGrader to score student responses on two open-ended questions about natural selection for six Key Concepts and three Naive Ideas (see Box 1). Two human graders (MJW and LSM) scored student responses for these same criteria. We resolved any disagreement among the humans by discussion, resulting in a consensus human score.

Statistical analysis

We measured inter-rater reliability (IRR) between the EvoGrader scores and the consensus human scores for each question, as outlined in (Hallgren, 2012). Because we were interested in the IRR of specific questions, we combined both pre-and post-instruction responses into a combined data set. We computed IRR both for each question as a whole, and separately for the key concepts and the naive ideas within each question. We chose to not compute IRR for each individual concept, or separately for pre- and post-instruction questions, because of the lower statistical power from examining each set separately, and the increase in multiple comparisons this would necessitate. We also compared the EvoGrader and human consensus scores by way of 2-tailed paired t-tests to test for differences in the number of key concepts or naive ideas scored. We conducted all statistical testing in R version 3.2.3 (R Core Team, 2013).

Results and Discussion

The Inter-Rater Reliability (IRR) of EvoGrader and the consensus human scoring of these questions is good, with values of 0.63 for the antibiotic resistance question and 0.55 for the mushroom question (Fig. 2). This means that more than half of the total variance in scoring across these 9 concepts is shared among the raters. Landis and Koch (1977) suggest that IRR values from Cohens kappa in the range of 0.6 to 0.8 indicate substantial agreement among coders, and values between 0.4 and 0.6 indicate moderate agreement (Landis and Koch, 1977). By these criteria, when all of the concepts are analyzed together, the IRR for the antibiotic question is strong, and the IRR for the mushroom question is moderate.

We further examined IRR separately for Key Concepts and Naive Ideas (Fig. 3), to examine whether there was a systematic difference between the two concept types. In the antibiotic resistance question, the IRR is notably higher for the Key Concepts than the Naive Ideas (0.63 v 0.17). In fact, the 95% confidence interval for the Naive Ideas IRR overlaps 0, meaning that the IRR is not statistically significantly different from ratings being assigned at random. Conversely, IRR in the mushroom question is consistent across the Key Concepts and Naive Ideas (0.51 and 0.55, respectively), showing no meaningful difference across concept type.

What can account for these differences in IRR? One thing to take note of is that when there is very low variation in a given raters scoring across responses, there is very little statistical power to detect shared variance across raters. As a thought experiment, imagine that two different raters assign scores of Yes to 10% of responses, and No to 90%. Even if the two raters both assigned their scores randomly, the two raters would be expected to agree 82% of the time. IRR analyses take into account the expected frequency of scoring agreement, but a low variance across responses for a given rater will negatively affect the statistical power of IRR analyses. This is reflected in the wide confidence intervals for the Naive Ideas in particular. For one, there are fewer potential Naive Ideas scored (since there are at most three Naive Ideas per student response, while at most six Key Concepts per student response). This skew in responses had a larger impact on the Naive Ideas in the antibiotic resistance question than elsewhere; EvoGrader only scored the entire class as expressing five total Naive Ideas in the antibiotic question; the consensus human score was 90. This is part of a general trend: for both questions, the human consensus score differed from the EvoGrader score, and by a statistically significant margin even when correcting for multiple comparisons (see Table 1; all adjusted p-values <0.05). For both questions, the human consensus score detected more Naive Ideas than EvoGrader did. However, the humans detected more Key Concepts than EvoGrader did for the antibiotic question (question 1), but fewer in the mushroom question (question 2).

Several factors may serve to lower the IRR from ideal levels. One obvious cause is mentioned in Box 1: some student responses demonstrate reasoning about natural selection, but do not answer the question asked. In these cases, the humans did not credit the student with any of the Key Concepts that did not address the question asked. EvoGrader, on the other hand, did not have this screening mechanism. Further, we analyzed both pre- and post-instruction responses jointly, and we expect the number of Naive Ideas expressed to decrease through instruction while we expect the number of Key Concepts expressed to increase through instruction. Such instructional effects would be a positive outcome for students, but both may reduce variance in the post-instructional scoring, reducing the statistical power to detect shared variance.

What can account for the difference in results between the two questions? There are two potentially salient contextual differences between the questions. One, the first question is a gain of a trait, while the second is a loss of a trait. Two, the two questions use different taxonomic groups as their examples. Both of these differences have been shown in the literature to be important to student reasoning (Nehm and

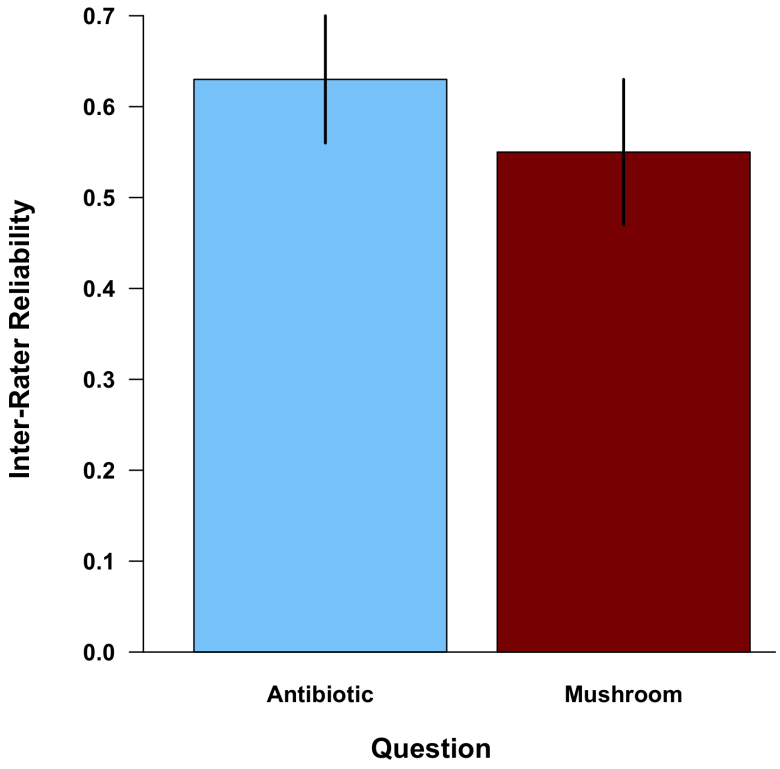

Figure 2: Inter-rater Reliability for Questions 1 and 2. Key Concepts and Naive Ideas are pooled within each question. Plotted values are Cohen's kappa. Error bars shown are 95% confidence intervals.

Comparison	t	df	p	adj. p
Antibiotic KC	5.779	67	$2.14 * 10^{-7}$	$8.58 * 10^{-7}$
Antibiotic NI	2.604	67	0.0113	0.0453
Mushroom KC	-2.806	71	0.00647	0.0259
Mushroom NI	3.384	71	0.00117	0.00466

Table 1: 2-tailed paired t-tests comparing EvoGrader and human consensus scoring of Key Concepts (KC) and Naive Ideas (NI). Negative values indicate more of these concepts detected by EvoGrader; positive values indicate more of these concepts detected by humans. A Bonferroni correction was used to generate the adj. p values.

Ha, 2011). In a future study, we will be able to disentangle these factors through a multifactorial design that considers multiple taxonomic groups and asks both a gain of trait and a loss of trait question within each.

Conclusions

EvoGrader is a useful tool for assessing student reasoning about natural selection. Even on questions not included in the training, it provides a reasonable level of reliability in scoring student responses on open-ended questions of a similar style to the ACORNS assessment. However, it is not

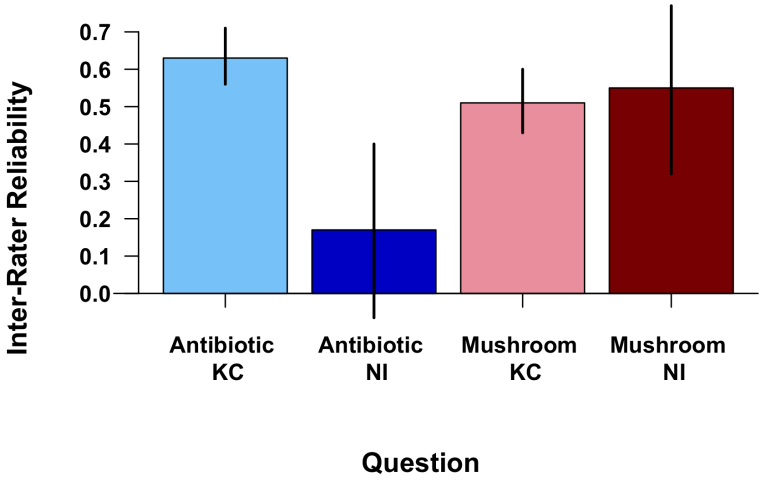

Figure 3: Inter-rater Reliability for Questions 1 and 2, broken down between Key Concepts (KC) and Naive Ideas (NI). Plotted values are Cohen's kappa. Error bars shown are 95% confidence intervals.

foolproof. In our study, EvoGrader credited students as displaying more Key Concepts, and fewer Naive Ideas, than our human raters did. In particular, EvoGrader may inaccurately credit student responses that do not address the specific question asked for evolutionary reasoning. For formative assessments, it can be a valuable tool to get a sense of student responses in a short period of time, but we caution against using EvoGrader to assign points to students, given its current limitations.

Acknowledgments. We thank Rohan Maddamsetti, Emily Dolson, Alex Lalejini, Anya Vostinar, Joshua Nahum, Brian Goldman, and Charles Ofria for helpful discussion during manuscript preparation. This work was supported by the National Science Foundation IUSE No. 1432563 and under Cooperative Agreement No. DBI-0939454. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Anderson, D. L., Fisher, K. M., and Norman, G. J. (2002). Development and evaluation of the conceptual inventory of natural selection. *Journal of Research in Science Teaching*, 39(10):952–978.

Andrews, T. M., Price, R. M., Mead, L. S., McElhinny, T. L., Thanukos, A., Perez, K. E., Herreid, C. F., Terry, D. R., and Lemons, P. P. (2012). Biology Undergraduates' Misconceptions about Genetic Drift. *CBE-Life Sciences Education*, 11(3):248–259.

Bishop, B. A. and Anderson, C. W. (1990). Student conceptions of natural selection and its role in evolution. *Journal of Research in Science Teaching*, 27.

Brewer, C. A. and Smith, D. (2011). Vision and change in undergraduate biology education: a call to action. *American Association for the Advancement of Science, Washington, DC*.

Butler, A. C., Marsh, E. J., Slavinsky, J. P., and Baraniuk, R. G. (2014). Integrating Cognitive Science and Technology Improves Learning in a STEM Classroom. *Educational Psychology Review*, 26(2):331–340.

Hallgren, K. A. (2012). Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in quantitative methods for psychology*, 8(1):23–34.

Hiatt, A., Davis, G. K., Trujillo, C., Terry, M., French, D. P., Price, R. M., and Perez, K. E. (2013). Getting to Evo-Devo: Concepts and Challenges for Students Learning Evolutionary Developmental Biology. *CBE-Life Sciences Education*, 12(3):494–508.

Kidziński, Ł., Giannakos, M., Sampson, D. G., and Dillenbourg, P. (2016). A Tutorial on Machine Learning in Educational Science. In Li, Y., Chang, M., Kravcik, M., Popescu, E., Huang, R., Kinshuk, and Chen, N.-S., editors, *State-of-the-Art and Future Directions of Smart Learning*, pages 453–459. Springer Singapore, Singapore.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33.

Mayfield, E. and Rosé, C. P. (2013). Open Source Machine Learning for Text. *Handbook of automated essay evaluation: Current applications and new directions*.

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

Nehm, R. H., Beggrow, E. P., Opfer, J. E., and Ha, M. (2012a). Reasoning about natural selection: diagnosing contextual competency using the ACORNS instrument. *The American Biology Teacher*, 74.

Nehm, R. H. and Ha, M. (2011). Item feature effects in evolution assessment. *Journal of Research in Science Teaching*, 48.

Nehm, R. H., Ha, M., and Mayfield, E. (2012b). Transforming biology assessment with machine learning: automated scoring of written evolutionary explanations. *Journal of Science Education and Technology*, 21.

Nehm, R. H. and Schonfeld, I. S. (2008). Measuring knowledge of natural selection: a comparison of the CINS, an open-response instrument, and an oral interview. *Journal of Research in Science Teaching*, 45.

R Core Team (2013). R: A language and environment for statistical computing.

Rector, M. A., Nehm, R. H., and Pearl, D. (2012). Learning the Language of Evolution: Lexical Ambiguity and Word Meaning in Student Explanations. *Research in Science Education*, 43(3):1107–1133.

States, N. L. (1900). *Next generation science standards: For states, by states*. National Academies Press.

Yahya, A. A., Osman, A., Taleb, A., and Alattab, A. A. (2013). Analyzing the Cognitive Level of Classroom Questions Using Machine Learning Techniques. *The 9th International Conference on Cognitive Science*, 97:587–595.