

## IMPROVING STUDENT LEARNING AND INSTRUCTIONAL EFFECTIVENESS THROUGH THE INNOVATIVE USE OF AUTOMATED ANALYSIS OF FORMATIVE ASSESSMENTS

Alexander Lyford<sup>1</sup> and Jennifer J Kaplan<sup>2</sup>  
<sup>1</sup>Middlebury College, Middlebury, VT, USA  
<sup>2</sup>University of Georgia, Athens, GA, USA  
 alyford@middlebury.edu

*Twenty years ago, Gal and Garfield (1997) argued that complex curricular goals associated with statistics learning could not be addressed adequately using solely multiple choice or short answer questions. In addition, the GAISE College Reports of 2005 and 2016 suggest that instructors use formative assessments, such as constructed-response questions, to improve student learning. In this paper, we demonstrate an innovative, forward-looking method for meeting the historical goals of improving student learning and instructional effectiveness through the use of algorithmic, automated categorization of student constructed responses to formative assessment items. We then demonstrate how instructors can use reports generated from their students' responses by the automated algorithms in an online portal. This portal allows instructors to integrate formative assessments into instruction quickly, especially in large-lecture classes.*

### INTRODUCTION

Assessing student conceptual understanding of statistics is a difficult task (Gal & Garfield, 1997). Gal and Garfield (1997) also argued that multiple choice and short answer questions may provide an incomplete assessment of this knowledge. Constructed-response (CR) questions, those in which students must answer using their own words, often provide a more insightful look into students' thought processes (Kuechler & Simkin, 2010). Unfortunately, developing effective constructed-response questions is difficult, and analyzing students' responses to these questions is often prohibitively time-consuming, particularly in high-enrollment, large-lecture format classes. This paper describes a project that seeks to bridge this gap by providing instructors with CR formative assessment items for pre-, post-, and mid-lecture use that are available for use via an online portal that provides real-time feedback to instructors about their students' responses. Through our free online portal, instructors are able to browse an array of statistics content questions. They can then select and download a question of interest. After distributing the selected question to their students, instructors can upload the responses to the portal and view an interactive report containing diverse feedback about their students' responses. These reports have been used successfully with university faculty in the biological sciences as part of long-term professional development in which the faculty modify their instruction and/or create new classroom activities to address the non-normative reasoning exhibited by their students. The feedback reports are not intended to be used for assigning grades to students; their function is to inform instructors of the progress students are making toward articulating normative reasoning.

In the paper we describe the process used to create the CR assessments and categorization models, describe the use of machine learning algorithms to categorize new student responses, and provide guidance for instructors on the process of uploading their students' responses to the online portal using a question in which students are asked to describe data shown as a histogram as the motivating example.

### QUESTION DEVELOPMENT CYCLE

The Automated Analysis of Constructed Responses (AACR; <http://create4stem.msu.edu/project/aacr>) research group has as a goal the creation of CR questions that can be scored using software to help us gain greater insight into student thinking about 'big ideas' in STEM. The process by which AACR questions and their associated categorization models are developed and validated is called the Question Development Cycle (QDC). In the first step of the QDC, *question writing*, a question is written and tested in a pilot study. If the question is found to elicit reasonable responses, the responses are moved to the next stage of the QDC, *exploratory analysis and rubric development*. If the responses are found to be problematic the

question undergoes revision and collection of more pilot data until the collected responses are deemed to be reasonable. In *exploratory analysis and rubric development*, many responses, typically between a few hundred and one thousand, are read and categories are created. Categorization rubrics can either be holistic or analytic (Airasian & Russell, 2008). In a holistic rubric, responses must belong to a single category in the rubric. In an analytic rubric, responses may belong to none, all, or some combination of the categories in the rubric. Some items in the portal utilize one of these types of rubrics, while others utilize both. The goal of any categorization rubric is to collapse the manifold responses into a few manageable categories. In addition, the creation of a categorization rubric is intended both for improving understanding of student responses and for preparing the responses to be categorized by machine learning algorithms, discussed in detail in the next section.

Several of the statistics content questions focus on students' understanding of graphs. In one such item, the Student Sleep question, students are asked to describe a unimodal, roughly-symmetric histogram showing the number of hours of sleep students in Georgia got on a given night. Given the nature of the Student Sleep prompt and typical of constructed-response questions in general, students often respond in diverse and unique ways—rarely are students' responses identical. After asking the Student Sleep question to over one thousand students at a large university in the U.S., an analytic rubric was constructed to categorize the responses (Table 1). Each of the responses was categorized according to its discussion of the graph's shape, center, and variability. Each response was also categorized as to whether or not the student answered within the context of the question.

Table 1: Student Sleep Categorization Rubric

Category	Requirements for belonging in each category
Shape	Students must correctly discuss the shape of the histogram by describing Student Sleep as symmetric, unimodal, bell-shaped, or approximately normal
Center	Students must give a valid measure of center (e.g., mean, median, mode, average) and correctly state its location.
Variability	Students must discuss either the range of the data, highlight potential outliers, locate the maximum and minimum values, or give an approximation of the variability directly (e.g. IQR, standard deviation).
Context	Students must answer the question within the context of the problem by using the appropriate variable with the appropriate units (e.g., 8 hours) and identifying the subject of each unit (e.g., students). At least two of the three following aspects: variable, units, and population, must be present

Once a rubric has been developed, the item moves to *hand coding* in which at least two independent raters code a training set of data. This process is typically done in stages, with the raters meeting to calibrate after coding about 100 responses and again after the entire training set is coded to resolve disagreements. The responses and coding are then sent to the machine learning algorithm ensembles for *model training*. Categorization rubrics, much like the one seen in Table 1, are the building blocks of the automated categorization methods—the machine learning algorithms. In essence, the machine learning algorithms described in the subsequent section attempt to mimic the categorization scheme employed in the rubric. Mimicking categorization schemes becomes difficult when rubric categories do not clearly distinguish responses belonging in the category from those that do not belong. Rubrics must therefore be explicit in their definitions of categories, and experts performing the hand-categorizing must be consistent in how they categorize similar responses. As such, the process of rubric creation is iterative and involves several stages of rubric refinement until sufficiently clear delineations are created. If the results of the model training indicate the ensemble results are valid and reliable, the QDC is finished and the question is loaded into the portal. Often, however, *model training* is a cyclic process in which hand coding, rubrics, questions, data processing, and/or algorithm tuning must be revised.

## MODEL TRAINING

Machine learning algorithms form a class of statistical algorithms that can learn from and make predictions about data (Kohavi & Provost, 1998). Several different machine learning algorithms are used by the portal to expedite the process of categorizing students' text responses, but all of the algorithms used are from a subset of algorithms known as supervised learning algorithms. Supervised learning algorithms require hand-categorized training data (Mohri et al., 2012). These algorithms use the training data to learn to make predictions about new, un-categorized responses. In essence, the machine learning algorithms attempt to mimic the categorization scheme of the training data coded by experts. Presently, there are eight supervised learning algorithms programmed into the portal: classification trees, bagging classification trees, boosting decision stumps, random forests, elastic-net regularized generalized linear models, maximum entropy modeling, scaled linear discriminant analysis, and support vector machines (see Kotsiantis et. al., 2007). Another common supervised learning algorithm, neural networks, was explored, but found to be less useful than the selected algorithms because of the relatively short length of the student responses.

To improve classification accuracy, the machine learning algorithms in the portal are used together in an ensemble. Ensemble learning is a technique that employs multiple machine learning algorithms in tandem, because the combined knowledge of multiple algorithms often produces more accurate results than any one algorithm alone (Dzeroski & Zenko, 2004). In this context, each algorithm in the ensemble makes a single classification prediction for any given student response. The votes from each algorithm are then weighted depending on that algorithm's performance on a set of testing data.

The ensemble vote weighting scheme differs from item to item. In general, based on training data, votes from algorithms with low rates of false positives are weighted higher when making a positive prediction, and votes from algorithms with high rates of false positives are weighted lower when making a positive prediction. The same pattern is true for negative predictions. In essence, each algorithm is weighted based on several criteria, most notably the rate at which its predictions are false positives or false negatives.

A unique ensemble of algorithms is created for each category of each assessment item [for further information on text classifiers, see Aggarwal & Zhai, 2012]. The performance of the ensemble relative to the category is measured by using 10-fold cross-validation. Thus, the hand-categorized data are split into ten groups of roughly equal size. Nine of these groups of data are combined to form a training set, and the remaining group becomes the test set. The ensemble is trained on the training set, and its performance is tested on the test set [see, for example, Kotsiantis et. al, 2007]. This process is repeated until each of the original ten groups of data is used as a test set. Metrics such as Cohen's Kappa, sensitivity, specificity, precision, recall, and overall accuracy are then used to evaluate the ensemble's performance. Trained ensembles with acceptable accuracy metrics are saved and stored on the web portal for use in predicting new, un-categorized student responses. Poorly performing ensembles are tuned to improve performance.

For the four categories associated with the Student Sleep question, shape, center, variability, and context (Table 1), the ensemble had 98.2%, 95.1%, 90.6%, and 99.1% agreement, respectively with the human categorizations. Values of Cohen's Kappa, a measure of agreement that takes into account the probability that an agreement between the ensemble and a human coder would happen by chance, larger than 0.8 are said to show strong agreement between two coders. The Kappa values for the categories in the Student Sleep question were 0.971, 0.946, 0.871, and 0.985, respectively. In general, the project aims to produce ensembles with agreement of at least 90% and Kappa values of at least 0.8.

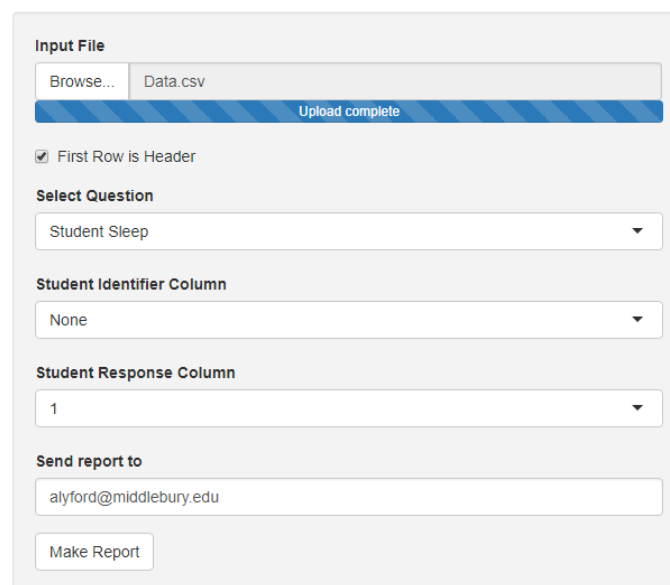
## THE PORTAL

The online portal contains more than one hundred ready-to-use questions spanning several STEM fields such as Biology, Chemistry, Genetics, and Statistics. Each of these questions is a constructed-response item, and the majority of questions require more than a sentence-length response. As more data are collected, additional questions will be added to the portal. A substantial amount of time has been spent by QDC researchers developing constructed-response questions and their corresponding rubrics for use in categorizing students' responses. For instructors, however,

the process of receiving feedback about their students' responses is significantly less time-consuming. There are four primary steps to this process of receiving feedback about student responses: 1. access the online database of ready-to-use, constructed-response questions about statistics, 2. select a question of interest and distribute it to students, 3. upload a file containing the students' responses to the online portal, and 4. explore the interactive feedback report provided by the portal

The online question database is accessed via the website <https://apps.beyondmultiplechoice.org/AutoReport/> and does not require a log in by the user. Questions can be filtered by discipline, such as chemistry or statistics, and by topic, such as central limit theorem or variability in graphs. After selecting a question of interest, instructors may then upload the question to their course management system. Instructors may then export their students' responses into an Excel or .csv file and upload this file to the web portal under the *Input File* drop-down menu. Instructors should select the corresponding question under the *Select Question* drop-down menu. If students' names or unique identifiers are attached to the data set, this should be indicated in the *Student Identifier Column* drop-down menu. Finally, the column with students' responses should be indicated in the *Student Response Column* drop-down menu.

After uploading the data and selecting the corresponding question, instructors may choose to view the interactive report in their web browser by clicking *Make Report*, or they may receive a link to the report via email by entering their email address in the *Send Report To* tab. The screenshot (Figure 1) shows an example in which a user has uploaded the file *Data.csv*, containing student responses in the first column, to the Student Sleep question. These data do not contain student identifiers, and the report will be visible online and via email.



The screenshot displays a web form titled "Input File". At the top, there is a file selection area with a "Browse..." button and the filename "Data.csv". Below this is a blue progress bar with the text "Upload complete". A checkbox labeled "First Row is Header" is checked. The "Select Question" dropdown menu is set to "Student Sleep". The "Student Identifier Column" dropdown menu is set to "None". The "Student Response Column" dropdown menu is set to "1". At the bottom, the "Send report to" field contains the email address "alyford@middlebury.edu", and a "Make Report" button is visible.

Figure 1. Portal User Interface

Reports created for fewer than one hundred student responses should take no longer than a few seconds to generate. Reports for several thousand student responses may take up to two minutes to generate. After making the report, the portal will direct users to the interactive feedback page. On this page, instructors can view the results of the machine learning algorithms' categorizations of their students' responses at a variety of levels. At a broad level, the report provides tables and graphs showing the proportion of responses falling into each of the possible rubric categories. For example, Figure 2 shows the results of one thousand student responses to the Student Sleep question. As evidenced by the graph, the majority of responses were categorized as containing discussions of shape and center, and these responses were often given in the context of the problem. Fewer than half of the student responses were categorized as containing a discussion of variability. This trend is seen throughout many questions asking students to describe a

histogram—students often identify the approximate center and overall shape of the graph, but rarely discuss the range, outlier points, maximum and minimum value, or other measures of variability.

There are numerous other interactive graphics available in the online report. Users can view web diagrams, which show the co-occurrences of words used in a single category. For example, responses in the Student Sleep question belonging to the context category often use both 'sleep' and 'hours' to describe the histogram. These words would have a high co-occurrence, and graphs of co-occurrences can help instructors better understand the types of responses categorized in a particular category.

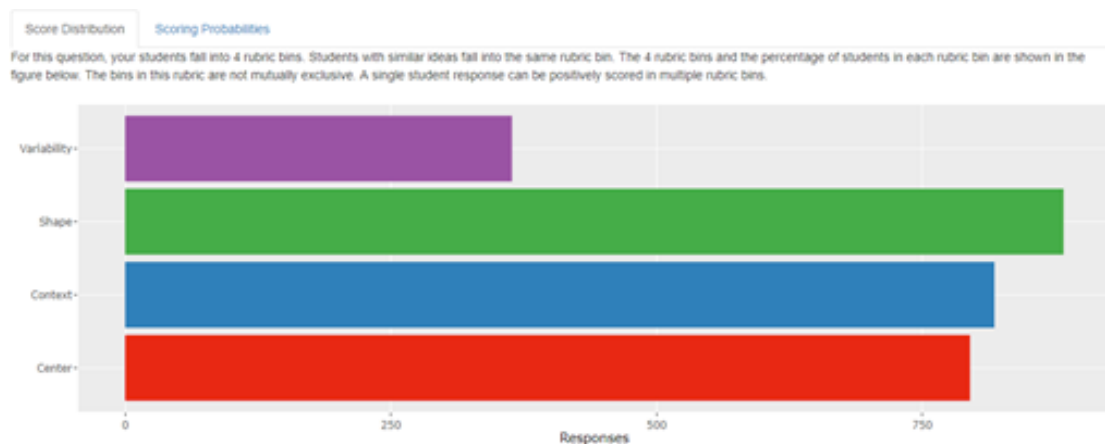


Figure 2. Summary of Response Categorizations

For analytic rubrics like the one used for Student Sleep, the portal also contains graphics about category-level co-occurrences. These graphs prompt users to select a category and then show users the proportion of responses in the selected category that belong to other categories—the category co-occurrences. This is particularly helpful for identifying archetype responses. For example, the co-occurrence rate of the center category to the variability category for Student Sleep is very low. That is, students that mention the center of the histogram are very unlikely to discuss the variability of the histogram. The co-occurrence rate of the variability category to the center category, however, is extremely high. Thus, students mentioning the variability of the histogram almost certainly also address the center of the histogram. These co-occurrences can provide additional insight into how students construct their responses.

In addition to aggregated output, the interactive report allows users to drill down to student-level response output. As part of this output, users can select responses that include a single categorization, responses that include a combination of categorizations, or responses that do not include a particular combination of categorizations. Figure 3 shows responses from the Student Sleep question that were categorized into the context category. Users can see how each response was categorized by the ensemble and the estimated probability that the ensemble made a correct categorization. Certain responses are more difficult for the ensemble to categorize due to the specific verbiage used in the responses. As such, some responses are categorized with a higher confidence than other responses. Tables and graphs like the one in Figure 3 help users better understand how the rubric and ensemble tandem categorized each student response.

## CONCLUSION

Integrating constructed-response assessments in everyday class activities is challenging. Analyzing the results of these assessments takes time and can rarely be done in time to provide meaningful formative feedback to the instructor. The innovative web portal described in this paper and its corresponding automated categorization processes can significantly expedite the rate at which CR assessment items are analyzed, allowing their use before, during, and after classroom activities. Through the use of ensembles of machine learning algorithms, students' responses are

categorized for use in modifying instruction into expertly-coded rubric categories in just a few seconds.

In the coming years, the portal will be expanded to include a wider variety of formative assessment items for statistics. These items will ultimately cover material taught in a standard introductory statistics course, although the assessment items themselves will be robust enough to be used in a variety of lower-level statistics courses. With the knowledge gained from the categorization of students' responses to these assessment items, instructors can better understand their students' knowledge, assess the effectiveness of in-class activities, and prepare for future classes.

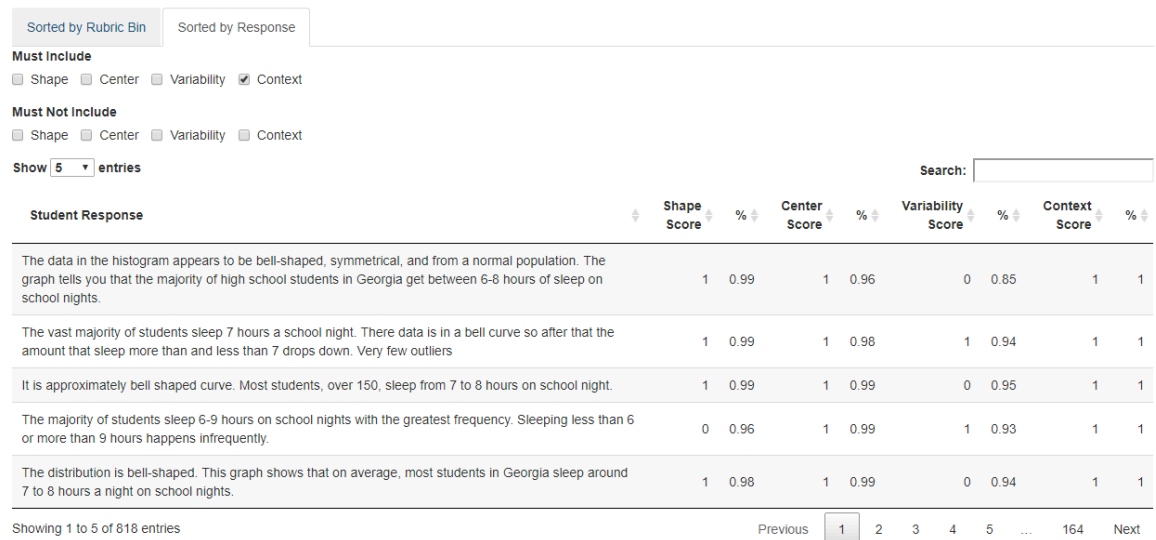


Figure 3. Detailed Response Categorization

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant No1322962. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the NSF.

REFERENCES

Aggarwal, C. C., & Zhai, C. (2012). A survey of text classification algorithms. In C.C. Aggarwal & C. Zhai (Eds.), *Mining text data* (pp. 163-222). Boston, MA: Springer.

Airasian, P. W., & Russell, M. K. (2008). *Classroom assessment: Concepts and applications (6th ed.)*. New York, NY: McGraw-Hill.

Dzeroski, S., & Zenko, B. (2004). Is combining classifiers with stacking better than selecting the best one? *Machine Learning, 54*(3), 255-273. Dordrecht, Netherlands: Kluwer Academic Publishers.

Gal, I., & Garfield, J. (1997). Curricular goals and assessment challenges in statistics education. In I. Gal & J. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 1-13). Amsterdam, Netherlands: IOS Press.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Informatica, 31*, 249-268.

Kohavi, R., & Provost, F. (1998). Glossary of terms. *Machine Learning, 30*(2), 271-274.

Kuechler W.L. & Simkin M.G. (2010). Why is performance on multiple-choice tests and constructed-response tests not more closely related? Theory and an empirical test. *Decision Sciences Journal of Innovative Education, 8*, 55-73.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of machine learning*. Cambridge, MA: MIT Press.