Labeling Mathematical Errors to Reveal Cognitive States

Thomas S. McTavish and Johann Ari Larusson

Center for Digital Data, Analytics, and Adaptive Learning Pearson, Austin, TX, USA
{tom.mctavish,johann.larusson}@pearson.com

Abstract. While technology can enhance learning, ironically, many online systems occlude learners' cognitive states because instructors do not directly observe students solving problems. In this paper, we show how we utilized an online mathematics homework system where students simply provided final answers to exercises. We then asked, "What can we infer about the cognitive state of the student if they gave an incorrect response?" Through data mining techniques, we found we were able to ascribe a particular type of mechanical error or misconception to 60-75% of the incorrect responses learners made on the subset of problems we analyzed. As such, we illustrate methods for extracting this data to discover knowledge components embedded in an exercise, expose item bias, and reveal learners' cognitive states.

1 Introduction

In education, "going digital" runs a gamut. The low end consists of digital forms useful for analytics, search, and archival. The high end consists of complete monitoring of learner interactions along with their biometric measurements [6]. While the high end of the spectrum offers the finest resolution of student learning, it also requires the learner to be, in a sense, "jacked in" and constantly surveilled, which may impede learning [9].

In this paper, we targeted the lower end of the digital spectrum by working with an online mathematics homework system where students simply entered a final, free text response to each exercise. We then asked, "What can we infer about the cognitive state of the student when they answered incorrectly?" Even without intermediate steps embedded in the task we were still able to label 60-75% of the errors in various exercises. We used relatively large sample sizes and exploited templates, each instance of the template having different parameters (e.g., "4+7" or "3+5"), which permitted us to see repeated patterns. We ascribed several types of errors to even seemingly straightforward exercises, demonstrating that several Knowledge Components (KCs) – the skills, concepts, and rules embedded in an activity [8] – may go into an exercise. Additionally, we were able to determine bias in the template that made the problem easier or more difficult depending on the parameters given. Methods such as ours illustrate approaches to inferring learners' cognitive states without complex assessment design and without perpetual interaction and surveillance.

C. Rensing et al. (Eds.): EC-TEL 2014, LNCS 8719, pp. 446–451, 2014.

2 Methods

Our data was comprised of user responses to online quizzes and tests using math exercises from a college level developmental math book. Responses were free text, final answers and graded as "correct" or "incorrect". The system randomly creates instances of each template using automatic item generation, often with certain constraints in the hopes of keeping the problem within the same domain and similar range of difficulty. In the two examples in this paper, one had 51 instances and the other had 1022.

We gathered student responses from each exercise. Since the response field is free text, it was straightforward to find those cases where several students converged to enter the same incorrect response. To determine *how* students arrived at their particular response, we looked across the instances of the template, comparing the peak responses to determine consistent patterns. After determining the type of error, we wrote mathematical expressions to match the specific error for the given input parameters of the instance. We therefore tagged those responses that had one or more errors attributed to them. (Since an incorrect response might match more than one error formula, the "Total inferred" row at the bottom of many of the tables we provide in our results is not a subtotal of each error category. Each tagged incorrect response was only counted once.) We then filtered out these cases and iteratively considered the remaining responses in attempts to further ascribe possible types of errors.

We determined bias in a template in the following three ways: We first considered the fraction correct with a particular instance as compared with the rest of the instances in a binomial test. We then looked at all instances that had a variable set to a specific value. Taking each variable across all of its values and performing the same binomial test allowed us to determine if and when a variable showed bias. We carried this one step further and performed the same binomial test on pairs of instance variables as well.

3 Results

3.1 "Find a Quotient" Example

Students were presented with the exercise "What is the quotient of x and 5?" where x was a uniformly random multiple of 5 in the range $[1000, 1250]$. We evaluated 2467 responses of which 379 were incorrect. With 51 possible instances with $x \in \{1000, 1005, \ldots, 1245, 1250\}$, there were only 7 incorrect responses per instance on average. Nevertheless, we noted that 31% of the errors followed the form $x \times 5$ and 12% of errors matched $x + 5$, indicating a misconception or misunderstanding surrounding "quotient". Interestingly, while many errors either multiplied or added, less than 1% of students gave a response that matched $x - 5$, implying that these students realized "quotient" did not involve subtraction.

We evaluated a similar exercise, simply stated as:

$$\text{Divide: } x\overline{)y}$$

(Type the whole number part and the remainder part of the quotient. Type 0 in the second answer box if there is no remainder.)

Users were provided with two text fields to type the whole number and remainder. This format and set of instructions with different conditions for x and y comprised many exercises of the preceding section in the book relative to the "Find a quotient" example. In this particular exercise x was an integer in the range $[2, 9]$ and y was an integer in the range $[1000, 5000]$, calculated such that there was no remainder, so users were always expected to give "0" as the remainder. The errors from this long division problem never followed the format of those errors just described in the "Find a quotient" example. For example, students never expressed the answer as $x \times y$. In fact, of those students who solved both exercises, 57% of those that got the "Find a quotient" example wrong had previously solved the long division problem correctly. Collectively, it implies that many students simply stumble on the definition of "quotient" or were thrown by the change in format to the question.

With the remaining responses to the "Find a quotient" example, we could see that if the correct answer contained a "0" digit, that many student responses omitted it. For example, if users were asked to find the quotient of 1015 and 5, which is 203, they might give "23". Such a response indicates a mechanical error or misconception surrounding place values and accounted for 10% of all errors. In fact, while the mean probability correct was 85% for this problem, when contrasting problems that had a "0" in their correct answer vs. those that did not, the probability of success was 81% ($p = 0.008$) and 87% ($p = 0.025$), respectively, indicating that the problem added another knowledge component when asking students to deliberately consider the place value. Collectively, we could characterize 58% of the errors from this problem, which are summarized in Table 1.

Table 1. Summary of errors to the "Find a quotient" example

ERROR	FREQ (%)
numerator × denominator	31
numerator + denominator	12
Omitted "0" in answer	10
Simply gave denominator	3
Simply gave numerator	2
numerator − denominator	1
Total inferred	58

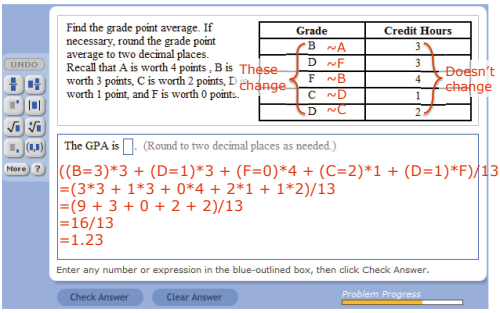

Fig. 1. Exercise to find the grade point average (GPA). Each row of grades in the left column can be an A, B, C, D, or F, except as specified. The credit hours column is fixed.

3.2 "Calculate a GPA" Example

As another example, we used the problem shown in Figure 1, which asks students to calculate a GPA. A nice feature of this template is that the *Credit Hours* column stays fixed. As such, all correct answers divide the sum of the grade points by 13. With 6322 exposures and a 43% success rate, we also had a large sample of errors ($n = 3634$) to work with. Since one grade letter was not permitted for each row, there were 46 possible instances such that correct answers will divide some numerator $\{3, 4, ..., 48, 49\}$ by 13 and round to the nearest hundredth. We contrasted the correct and incorrect answers. Interestingly, several incorrect responses were given as a number properly divided by 13 and rounded. In this case, learners obviously calculated an incorrect numerator. This accounted for 24% of the errors. Another interesting pattern we observed was that many incorrect responses ended in .0, .2, .4, .6, or .8. In other words, learners were dividing by 5, the number of classes, instead of 13, the number of credits. This misconception revealed itself in a few ways as enumerated in Table 2.

We also observed a lot of bias in this template. Unsurprisingly, if the GPA to be calculated was 2.0 or 3.0, the problem was significantly easier (50% fraction correct) than if it was $12/13 = 0.92$ (27% correct, and the most significant response). We also observed that if three "Fs" were presented, the fraction correct dropped to 31% ($p = 6 \times 10^{-5}$). While four "Fs" had a success rate of 27%, it was not statistically significant because of the few samples. We also found that when "As" were absent, the fraction correct was 39% ($p = 0.0016$) and when two "As" were given the success rate went up to 46% ($p = 0.016$). Collectively, these results hint that students understand the value of an "A" more than other letters, especially "F". Again, we speculate that perhaps students found it difficult to imagine taking a course and not receiving any credit for it.

4 Discussion

In this paper, we showed how incorrect, final responses to mathematics exercises can be aggregated and analyzed to infer specific types of errors. We were able to label 60-75% of the errors in the exercises we sampled and ascribe at least a half dozen error types to each problem. This work simultaneously quantifies intra-template variability, possible knowledge components in an exercise, and enables students' cognitive states to be inferred by labeling their errors.

4.1 Shortcomings and Promise of Technology

The most time-consuming task, while also being the most critical component of teaching, is the instructors' continual assessment of student performance [4]. Students are doing more and more online, so instructors do not directly observe students solving problems. In many cases, as in systems such as the one we used where students simply enter a final answer, instructors do not see worked solutions. Furthermore, as convenient as automatic scoring may be, many scoring

Table 2. Summary of errors for the GPA problem. "Summed grades" are where students applied $A = 4$, $B = 3$, $C = 2$, $D = 1$, and $F = 0$ to sum their set of grades, but did not multiply by the credit hours.

ERROR	FREQ (%)
Wrong numerator (Counting error)	24
Summed grades, but did not multiply by credit hours, then divided by number of classes instead of credit hours	21
Answer out of bounds > 4.0	13
Rounding errors	11
Summed credit hours, but did not multiply by any grades, then divided by the number of classes ($13/5 = 2.6$)	10
Summed grades and divided by the grade points	5
Correct numerator, but divided by number of classes instead of number of credit hours	5
Summed grades, but did not multiply by credit hours (subset of "Wrong numerator")	2
Did not credit "Fs"	1
Correct numerator, but never divided	1
Summed grades and never divided	1
Total inferred	75

systems do not incorporate subtleties. More often than not, an item is deemed correct or incorrect and scored by the item's weight.

Our approach offers a window into students' cognition that is often occluded by technology. We show ways of discriminating misconceptions from mechanical errors, which can then be communicated to instructors and students. The system itself can score items differently and adapt to the learner. The most critical component of successful learning is prompt, detailed, positive and timely feedback on student work [2]. Indeed, while technology may remove certain face-time with instructors and a given system my provide insufficient metrics and feedback to instructors and students, technology also has the capability of providing an informative, data-rich environment [3].

4.2 Bug Libraries Revisited

Analytics surrounding mathematical errors is hardly new. Our work is reminiscent in spirit to the "bug library" work performed by VanLehn and Brown [1,7]. They constructed libraries of incorrect repairs to impasses to simple mathematics problems. Their approach combined cognitive models to algorithmically generate possible bugs and extensive analysis on handwritten mathematics problems solved by students. They ascribed, for example, over 70 bugs for subtraction. As the complexity of mathematics problems grows, constraining the error space to a limited, computational domain is nearly impossible. Automated methods to infer bugs are still needed. Nevertheless, we demonstrate some analytic methods and considerations that might be utilized in the construction of automated

methods directed at complex problems. For example, the "calculate a GPA" exercise had a few important design characteristics. For one, it kept the *Credit Hours* column fixed. This made a discrete set of correct solutions by which the incorrect solutions could be contrasted. Furthermore, two of the variables – the number of classes and the total credits – were each a prime number. Therefore, the incorrect answers were, in a sense, more easily parsed. If the number of credits were 10, for example, then it might not have been as apparent that many students were dividing by the number of classes instead of the number of credits.

4.3 Knowledge Components and Bias

Knowledge components may be explicitly incorporated by task authors, or statistically revealed if unknown [5]. By labeling errors, we illuminate the KCs that have failed, demonstrating another means of discovering them. Template bias also indicates the KCs at play. We saw in our "Find a quotient" example that students often omitted a "0" in their answer if the correct solution included it. Such bias revealed a "0 in the tens place" KC that was absent in the other instances. Likewise, our GPA example showed more students struggling with "Fs", incorrectly rounding, or having problems with the division when the GPA was not a whole number. This revealed that different KCs were being challenged with different instances. Collectively, our work demonstrates means of capitalizing on digital approaches even when much work may be performed offline.

References

1. Brown, J.S., VanLehn, K.: Repair theory: A generative theory of bugs in procedural skills. Cognitive Science 4(4), 379–426 (1980), http://onlinelibrary.wiley.com/doi/10.1207/s15516709cog0404_3/abstract
2. Brown, S., Race, P., Rust, C.: Using and experiencing assessment. Assessment for Learning in Higher Education, 75–85 (1995)
3. Conole, G., Warburton, B.: A review of computer-assisted assessment. ALT-J 13(1), 17–31 (2005), http://journals.co-action.net/index.php/rlt/article/view/10970
4. Crooks, T.J.: The impact of classroom evaluation practices on students. Review of Educational Research 58(4), 438–481 (1988), http://rer.sagepub.com/content/58/4/438, doi:10.3102/00346543058004438
5. Junker, B.W., Sijtsma, K.: Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. Applied Psychological Measurement 25(3), 258–272 (2001), http://apm.sagepub.com/content/25/3/258, doi:10.1177/01466210122032064
6. Picard, R.W.: Affective Computing. MIT Press (2000)
7. VanLehn, K.: Mind Bugs: The Origins of Procedural Misconceptions. MIT Press (1990)
8. VanLehn, K.: The behavior of tutoring systems. International Journal of Artificial Intelligence in Education 16(3), 227–265 (2006), http://iospress.metapress.com/content/AL6R85MM7C6QF7DR
9. Wang, F., Hannafin, M.J.: Design-based research and technology-enhanced learning environments. Educational Technology Research and Development 53(4), 5–23 (2005), http://link.springer.com/article/10.1007/BF02504682