Discovering and Describing Types of Mathematical Errors

Thomas S. McTavish*
Center for Digital Data, Analytics & Adaptive Learning
Pearson
tom.mctavish@pearson.com

Johann Ari Larusson
Center for Digital Data, Analytics & Adaptive Learning
Pearson
johann.larusson@pearson.com

ABSTRACT

Given a large number of incorrect responses to mathematical exercises, we ask, "What errors might the learner have made to arrive at their answer?" Even though our data does not contain intermediate steps, we find that we are able to infer well over 50% and sometimes over 90% of the types of errors learners make on an exercise when they only supply final answers. Our approach capitalizes on the sheer volume of data to highlight patterns and the fact that these exercises come from item banks of mathematical templates. Since items generated from mathematical templates deliver different parameters to different learners (e.g., one learner might see $y = 2x + 3$ while another learner might see $y = 3x + 5$), misconceptions and mechanical errors are more easily recognized. We enumerated different errors for simpler-stated problems and utilized other forms of signal analysis in other cases to uncover error types. Our results show that there are many types of errors even for seemingly simple problems, and we can quantify their relative degrees of prevalence. We can also determine bias in the templates that make a problem easier or more difficult depending on which parameters are used. Since error categories correlate with knowledge components, our work highlights the relative degree of knowledge components embedded within a problem and exposes some knowledge components that may otherwise remain unconsidered.

Keywords

Error Analytics, Misconceptions, Knowledge Components, Template Bias, Automatic Item Generation

1. INTRODUCTION

The etymology of the word *demonstrate* comes from the latin *dē* ("concerning") + *mōnstrō* ("I show") [25]. The noun form of *mōnstrō* has become *monster* in English, the word for something that warns or instructs. Medicine and biology have a long heritage of learning from abnormal patients and aberrant individuals, these so-called "monsters". For example, in the 19th century, after treating patients who exhibited severe speech and language deficits, Drs. Broca and Wernicke proposed areas of the brain giving rise to speech and language after their postmortem analysis revealed brain injuries at specific sites [7, 24]. In similar spirit, MRI scanning of stroke victims today continues to reveal the functional map of the brain [4, 18]. Likewise, the field of genetics is largely built around animal models such as the mouse and fruit fly where standard practice is to "knock in" or "knock out" genes and observe the phenotypes of the mutants [5, 21].

Indeed, errors help frame "normal". In the context of learning, the types of errors that are revealed in a task demonstrate areas of confusion and the hurdles that need to be overcome to attain mastery. Errors therefore have a strong correspondence with the knowledge components (KCs) – the skills, concepts, and rules of a problem [12, 16, 17, 22]. It has been found that those who have achieved mastery categorize problems differently than those who are novices, as their categorization is more shallow [3]. Prior work has shown that more often than not, student errors in simplistic fraction multiplication word problems concern the vocabulary used rather than the actual math itself [8, 13]. Furthermore, evidence exists demonstrating not only a causal relationship between students' prerequisite knowledge, or lack thereof, and errors in problem solving, but also that gaps in prior knowledge negatively impacts students' accrued learning [23]. Also, strategic errors in instruction, have been shown to be at fault and contribute to particular limitations in prior knowledge [2]. Such research highlights that errors can help sculpt and define KCs. We will argue that in many cases, they may be two sides of the same coin.

In the work presented in this paper, we evaluated incorrect responses to mathematical exercises to determine and quantify specific types of errors learners made. Even though our data contained only final answers, we were still able to label 60-90% of the errors without intermediate step data. Our approach exploited the relatively large number of samples of each problem (hundreds or low thousands) and took advantage that the exercises came from templates, each instance of the template having different parameters (e.g., "$4 + 7$" or "$3 + 5$"). Comparing across template instances permitted us to see repeated patterns. Sometimes setting incorrect responses against the backdrop of correct solutions provided clearer interpretations of the incorrect response to more eas-

*Corresponding author.

ily label it. We found we could ascribe several types of errors to even seemingly straightforward exercises, demonstrating that several KCs 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. As such, bias illuminated the challenges delivered to some learners that were not given to others. While such bias has strong implications for assessments, it also permits us to dissect the exercise further and label a KC associated with some values of the template's parameters that is otherwise absent in other instances.

2. METHODS

2.1 Data

Our data consisted of student responses to online math problems from a college level developmental math book. Students, largely from the U.S., were enrolled in courses spanning Fall semester 2012 through 2013 that used the Pearson MathXL® homework system. All responses were from quizzes or tests. Students may have seen an exercise in a prior homework, or, somewhat rarely, may have taken a quiz or test multiple times to have multiple exposures to the exercise, but we did not factor this into our analysis. We used the final answer to the problem, which was a free text field. Therefore, learners could enter strings that could remain string literals, or be parsed into numbers. Alternatively, students could use an equation editor to enter mathematical expressions into the field. Responses were either labeled "correct" or "incorrect".

The system employs automatic item generation from math templates, randomly creating instances of the exercise, often with certain constraints in the hopes of keeping the problem within the same domain and similar range of difficulty. By the combinatoric nature of the parameters used in some templates of the system, some exercises have nearly an infinite number of possible instances. For this study, we concentrated on some templates that had few instances (3 - 24) for all but the "GPA" example, which had 1022 instances.

2.2 Our approach

We collected student responses to each instance of each exercise. For the cases with few instances, we looked at the distribution of incorrect responses. Our null hypothesis is that incorrect responses are random guesses. Since the response field is free text, this means that the distribution of possible answers will be very large under the null hypothesis. It was straightforward, then, to find those cases where several students converged to enter the same incorrect response. What was not so apparent, and remains the bottleneck of our approach, was determining *how* students arrived at their particular response. For this, we looked across the instances of the template, comparing the peak responses to determine consistent patterns. This is perhaps best illustrated with an example as shown in Table 1 that shows a sampling of responses where four or more students gave the same response to the question "Find the reciprocal of the number x." Looking across the instances where x is 2, 3, 4, or 5, there are repeated patterns of users giving the negative of the number, or framing the response as $x/1$ instead of the correct response of $1/x$. The table also shows many students simply echoing the number given. Because these patterns repeat themselves across the different instances, we can more confidently define the type of error by seeing how it generalizes. 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. (Because 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.

Our "GPA" example had a different output from the others. In this case, only a handful of students saw the same instance. We therefore contrasted the distributions of correct and incorrect responses through visualizations as described in Section 3.6.

We determined bias in a template in the following ways: Firstly, we considered the fraction correct from students that saw a particular instance as compared with the rest of the instances in a binomial test. This permitted us to see if a set of specific instances were easier or harder. 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 illustrated in Figure 1.

3. RESULTS

We applied our method across several different templates as highlighted in each subsection.

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. About 5% of error responses simply echoed the numerator, x, or the denominator, "5". With the remaining responses, 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,