Chapter 17: Data Mining Large-Scale Formative Writing

Peter W. Foltz[1,2], Mark Rosenstein[2]

[1] Institute of Cognitive Science, University of Colorado, USA
[2]Advanced Computing and Data Science Laboratory, Pearson, USA
DOI: 10.18608/hla17.017

ABSTRACT

Student writing in digital educational environments can provide a wealth of information about the processes involved in learning to write as well as evidence for the impact of the digital environment on those processes. Developing writing skills is highly dependent on students having opportunities to practice, most particularly when they are supported with frequent feedback and are taught strategies for planning, revising, and editing their compositions. Formative systems incorporating automated writing scoring provide the opportunities for students to write, receive feedback, and then revise essays in a timely iterative cycle. This chapter provides an analysis of a large-scale formative writing system using over a million student essays written in response to several hundred pre-defined prompts used to improve educational outcomes, better understand the role of feedback in writing, drive improvements in formative technology, and design better kinds of feedback and scaffolding to support students in the writing process.

Keywords: Writing, formative feedback, automated scoring, mixed effects modelling, visualization, writing analytics

Writing is an integral part of educational practice, where it serves both as a means to train students how to express knowledge and skills as well as to help improve their knowledge. It is well established that in order to become a good writer, students need a lot of practice. However, just practicing writing is insufficient to become a good writer; receiving timely feedback is critical (e.g., Black & William, 1998; Hattie & Timperley, 2007; Shute, 2008). Studies of formative writing in the classroom (e.g., Graham, Harris, & Hebert, 2011; Graham & Hebert, 2010; Graham & Perin, 2007) have shown that supporting students with feedback and providing instruction in strategies for planning, revising, and editing their compositions can have strong effects on improving student writing.

Text as Data

Writing is a complex activity and can be considered a form of performance-based learning and assessment, in that students are performing a task similar to what they will typically be expected to carry out in their future academic and work life. As such, writing provides a rich source of data about student content knowledge, expressive skills, and language ability. Thus, writing affords making multiple inferences about the nature of student performance based on the textual information.

Currently most writing is mediated by computer, which provides an opportunity to study and impact writing learning at a depth and over time periods that were just not practical with paper-based media. For instance, Walvoord and McCarthy (1990), with a series of collaborators, conducted classroom studies over nearly a decade, gathering artifacts such as student journals, drafts, and final papers to build understandings of writing instruction. Much of the effort to conduct the study was in the collection and hand analyses. Today, with computer-based writing, such resources are more readily available as part of the writing process, and are in a form where natural language processing and machine learning can be automatically employed. By applying appropriate learning analytic methods, textual information can therefore be automatically converted to data to support inferences about student performance.

Automated analyses have been applied to understanding aspects of writing since the 1960s. Content analysis (e.g., Gerbner, Holsti, Krippendorff, Paisley, & Stone, 1969; Krippendorff & Bock, 2009) was designed to allow analysis of textual data in order to make replicable, valid inferences about the content. However, the methods focused primarily on counts of key terms used in the texts. Ellis Page (1967) pioneered techniques to convert the language features of student writing into scores that correlated highly with teacher ratings of the essays. With the advent of increasingly more sophisticated natural language processing and machine learning techniques over the past 50 years, automated essay scoring (AES) has now become a widely used set of approaches that can provide scores and feedback instantly. Research on AES systems has shown that their scoring can be as accurate as human scorers (e.g., Burstein, Chodorow, & Leacock, 2004; Landauer, Laham, & Foltz, 2001; Shermis & Hamner, 2012), can score multiple traits of writing (e.g., Foltz, Streeter, Lochbaum, & Landauer, 2013), and can be used for feedback on content (e.g., Foltz, Gilliam, & Kendall, 2000; Foltz et al., 2013).

While much of the focus in the evaluation of AES has examined the accuracy of the scoring and the different types of essays that can be scored, AES also has wide applicability to formative writing, where evaluation can focus more on how it aids student learning. Human assessment of writing can be time consuming and subjective, limiting the opportunities for students to receive feedback. As a component of a formative tool, AES can provide instantaneous feedback to students and support the teaching of writing strategies based on detecting the types of difficulties students encounter. For example, when incorporated into classroom instruction, students are able to write, submit, receive feedback, and revise essays multiple times over a class period. All student writing is performed electronically, and is automatically scored and recorded, providing a record of all student actions and all feedback they received. This archive permits continuous monitoring of performance changes in individuals as well as across larger groups of students, such as classes or schools. Teachers can analyze the progress of each student in a class and intervene when needed. In addition, it now becomes possible to chart progress across the class in order to measure effectiveness of curricula and teaching strategies as reflected in student writing performance. A number of formative writing tools using automated scoring have been developed and are in use, including WriteToLearn™ (W2L; Landauer, Lochbaum, & Dooley, 2009), Criterion, (Burstein, Chodorow, & Leacock, 2004), OpenEssayist (Whitelock, Field, Pulman, Richardson, & Van Labeke, 2013), and Writing Pal (Roscoe & McNamara, 2013).

Data Mining Applied to Writing

Automated formative assessment of writing provides a rich data set to examine the changes in writing performance and system features that influence that performance. With the increasing adoption of digital educational environments, there are new opportunities to leverage the data from student interactions in these environments as evidence (e.g., DiCerbo & Behrens, 2012). Recent work has begun to extract and analyze writing from writing assignments, from peer grading exercises, as well as from collaborative forum discussions in order to examine student performance. While there have been a number of overviews of data mining methods (e.g., Peña-Ayala, 2014; Romero & Ventura, 2007; Romero & Ventura, 2013), there has still been little focus on large-scale data mining of formative writing. With the advent of more powerful computational discourse tools, new techniques are emerging (e.g., Buckingham-Shum, 2013; McNamara, Allen, Crossley, Dascalu, & Perret, this volume; Rosé, this volume).

Some studies have examined large corpora of student writing, although not focused on the aspects of formative feedback. For example, Parr (2010) analyzed 20,000 essays written to 60 different prompts at different grade levels in order to measure how writing skills develop for different genres of essays. All scoring of the essays was performed by human scorers, although tools were provided to make the scoring easier and to ensure consistency. Deane and Quinlan (2010) performed analyses using the e-Rater automated scoring engine to extract features from thousands of essays and then performed factor analysis in order to examine developmental levels and linguistic dimensions of writing. Deane (2014) also used automated scoring of essays from a multi-state implementation, analyzing features from keystroke logs and the essays themselves, in order to predict factors of writing ability and reading level.

Aspects of the formative process have also been examined using smaller samples of data; for example, the research on collaborative writing at the University of Sidney (Calvo, O'Rourke, Jones, Yacef, & Reimann, 2011; Reimann, Calvo, Yacef, & Southavilay, 2010) used student log data and automated assessment to support writing. In their work, they analyzed grammatical and topical aspects of writing as well as log files of the sequences of revisions and writing activities in order to understand team writing processes. In addition, research has performed fine-grained analysis of writing by coupling log data with physiological monitoring, such as eye tracking (e.g., Leijten & Van Waes, 2013). WhiteLock et al. (2013, 2015) used visualizations of textual features of essays, including displays of key words and phrases and information about essay structure across multiple essays as a way to allow

students and instructors to understand aspects of the content of the essays. These visualizations can then be used as the basis for providing advice for improving student writing.

Other research involving writing and data mining has considered writing as a secondary task, such as Crossley et al. (2015), who examined student writing in discussion forums within MOOCs to predict whether a student would successfully complete the course, and White and Larusson (2014), who developed lexical analysis techniques to analyze changes in student writing to detect when students reach the point when they sufficiently understand a core concept in order to re-express it in their own words. Finally, analyses of feedback during the revision process in online systems (e.g., Baikadi, Schunn, & Ashley, 2015; Calvo, Aditomo, Southavilary, & Yacef, 2012) has shown what kinds of feedback can be most effective in the revision process. The majority of these studies focused on analyses based on tens to hundreds of students, so while they inform the use of data mining techniques and provide critical information on the role of formative feedback, they have not yet been scaled larger administrations.

This chapter builds on the above approaches to describe an approach to large-scale analysis of writing by applying data mining to components of the formative writing process on hundreds of thousands to over a million samples of writing collected from a formative online writing system. The analyses are used to investigate specific classes of questions about how a formative system is currently being used, its efficacy, and how understanding current use yields suggestions for improved learning, both through improving the system implementation and by introducing direct interventions aimed at students using the system. The chapter illustrates approaches utilizing descriptive statistics of performance as well as formally modelling changes in performance. While the chapter focuses on methodology, the intent is to illustrate how writing data can be used more generally to inform decisions about the quality of student learning, about the effectiveness of implementation in the classroom, as well as the effectiveness of the digital environment itself as an educational tool.

ONLINE FORMATIVE WRITING SYSTEM

The context used to illustrate the power of data mining in the lifecycle of a large-scale implementation was conducted with student interaction data from the formative writing assessment system WriteToLearn™. WriteToLearn™ is a web-based writing environment that provides students with exercises to write responses to narrative, expository, descriptive, and persuasive prompts as well as to read and write summaries of texts in order to build reading comprehension. Students use the software as an iterative writing tool in which they write, receive feedback, and then revise and resubmit their improved essays. The automated feedback provides an overall score and individual trait scores such as "ideas, organization, conventions, word

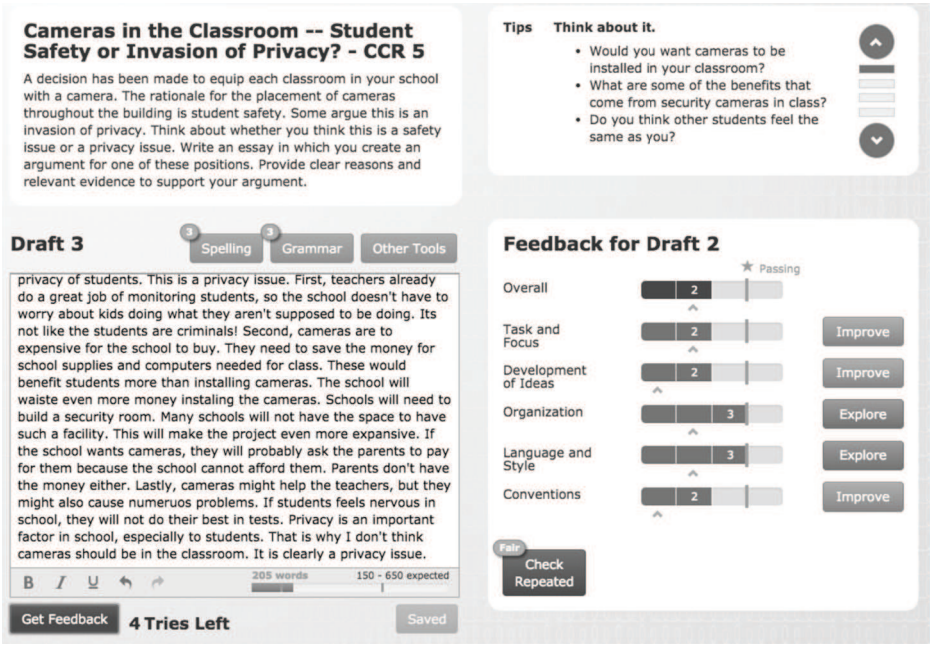

Figure 17.1. Essay Feedback Scoreboard. WriteToLearn™ feedback with an overall score, scores on six popular traits of writing, as well as support for the writing process.

choice, and sentence fluency." The student can also view supplemental educational material to help them understand the feedback, as well as suggest approaches to improve their writing. In addition, grammar and spelling errors are flagged. Figure 1 shows a portion of the system's interface, in this case illustrating the scoring feedback resulting from a submission to a 12th grade persuasive prompt. Evaluations of WriteToLearn™ have shown significantly better reading comprehension and writing skills resulting from two weeks of use (Landauer et al., 2009) as well as validating the system scores being as reliable as human raters, and significantly improved end-of-year pass rates on a statewide writing assessment (Mollette & Harmon, 2015).

Algorithm for Scoring Writing

WriteToLearn's™ automated scoring is based on an implementation of the Intelligent Essay Assessor (IEA). IEA is trained to associate extracted features from each essay to scores assigned by human scorers. A machine-learning-based approach is used to determine the optimal set of features and the weights for each of the features to best model the scores for each essay. From these comparisons, a prompt- and trait-specific scoring model is derived to predict the scores that the same scorers would assign to new responses. Based on this scoring model, new essays can be immediately scored by analysis of the features weighted according to the scoring model. The focus in this chapter is not on the actual algorithms or features that make up the scoring, as those have been described in detail elsewhere (see Landauer et al., 2001; Foltz et al., 2013). Instead, the focus is how the trail left by automated scoring and student actions can be used to monitor learning across large sets of writing data and facilitate improvements in the formative system.

Data

The data comprised two large samples of student interactions with WriteToLearn™ collected from U.S. adoptions of the software. One set comprised approximately 1.3 million essays from 360,000 assignments written by 94,000 students collected over a 4-year period. The second set represented approximately 62,000 student sessions with nearly 900,000 actions. The data included student essays and a time-stamped log of all student actions, revisions, and feedback given by the system. Essays were recorded each time a student submitted or saved an essay, resulting in a record of each draft submitted. The essays were written to approximately 200 pre-defined prompts. No human scoring was performed on these essays. All essay scores were generated by automated scoring, with the prediction performance of the models validated against human agreement from test sets or using cross-validation.

Analyses Enabled by the Approach

At all stages in the lifecycle of a formative system – design, implementation, deployment, redesign, and maintenance – analysis of actual use via analytics applied to log data can inform improvements to the system. As Mislevy, Behrens, Dicerbo, and Levy (2012) note, there is interplay between evidence-centred design, which represents best practices when a system is first conceived and data mining student actions of the implementation that reflects actual use, where each is critical in building and evolving educational systems. From the design phase, we are interested in analyzing use data to assess our assumptions and, in our case, determining if cycles of writing, feedback, and revising improves writing performance and at what rate and whether the rate of improvement differs among the traits of writing. In terms of pedagogical theory, we want to understand what mix of writing, mechanics feedback, content feedback, and revising leads to optimal learning, and potentially how to individualize advice to students and teachers. Currently the system by default allows six revision/feedback cycles with teachers able to customize the limit, and use data should help develop guidelines for this feature. Another quite productive form of analysis is to model student performance; here we discuss a mixed effects model that allows us to estimate the relative difficulty of the prompts. Prompts typically are assigned a grade level when developed, but modelling allows us to determine if the prompt is correctly labelled; using performance data from millions of essays written to a prompt allows finer grained levelling.

Many additional types of analysis are possible with writing-log data than there is room to detail in this chapter (see also Calvo et al., 2012; Deane, 2014). Two areas we have found particularly promising are evaluation of teachers' instructional strategies; for instance, in terms of which prompts were chosen and how long (a single class period, a week, longer) students were allowed to write to a prompt. While systems such as the one described here have professional development instruction for teachers as well as teachers' guides, it is astonishingly useful to observe how the system is actually used in classrooms in order to uncover new strategies and measure the relative effectiveness among the strategies. Another area that we lack space to describe in detail is fine-grained analysis of student actions. For instance, it is possible to tell when and where in the writing process a student exploits a help facility, and often possible to infer when a student should have taken advantage of a facility but didn't – from which it may be possible to infer redesign choices in terms of user interface layout and other design issues. Additional discussion of some of these directions can be found in Foltz and Rosenstein

(2013), Foltz and Rosenstein (2015), and illustrated in Foltz and Rosenstein (2016).

VALIDATING THEORY

Does Writing and Revising Result in Improved Writing Performance?

Formative writing systems are designed to support a rapid cycle of *write*, *submit*, *receive feedback*, and *revise*. This cycle is one of the key differentiators of automated formative writing from standard classroom writing practice, where human scoring of essays is time consuming so students cannot receive immediate feedback. Thus, it is critical to determine how often students submit and revise essays and determine the factors and time paths that lead to greatest success. This can help address questions of whether revising results in better writing, as measured by the automated scores and what patterns of use facilitate the most rapid improvement.

Using a subset of the data, we examined writing over a single semester in which teachers in three grades (5th, 7th, and 10th) across an entire state assigned writing exercises to students. During that period, 21,137 students wrote to 72,051 assignments (an average of almost four assignments per student) with 107 different unique writing prompts assigned. These assignments resulted in 255,741 essays submitted and scored over the period of analysis. For each submission, students received feedback and scores on their overall essay quality, as well on six different writing traits: ideas, organization, conventions, word choice, sentence fluency, and voice. While there was a wide distribution in the number of revisions students made, most students made more than one revision, with most making up to five revisions. Figure 2 shows the score improvement (score on last attempt minus score on first attempt) for students who wrote multiple drafts. It shows improvement for each of the six writing traits as well as the overall score. There is a clear trend indicating that more revisions equal higher scores. With the typical five revisions, the average student score improved by almost one score point (out of a maximum of 6). Generally, we see greatest improvement in scores for content-based features, such as ideas, voice, and organization, and less for features related to writing skill, such sentence fluency and writing conventions. The smoothness of the curves and small error bars are due to the large number of data points for each revision from 0 to 5.

Time Spent Between Revisions

We can further investigate the impact on student performance of the time-spent writing before requesting feedback to better understand the best allocation of time among the *write*, *submit*, *feedback*, and *revise* phases. Using data from approximately 1.1 million student writing attempts across a wide range of users of WriteToLearn™, we calculated the change in student grade (e.g., improvement from one draft to the next) based on how much time was spent between drafts. The change in grade shown in Figure 3 indicates that the improvement in writing score generally increases up to about 25 minutes at which point it levels off and begins to drop. In addition, most of the nega-

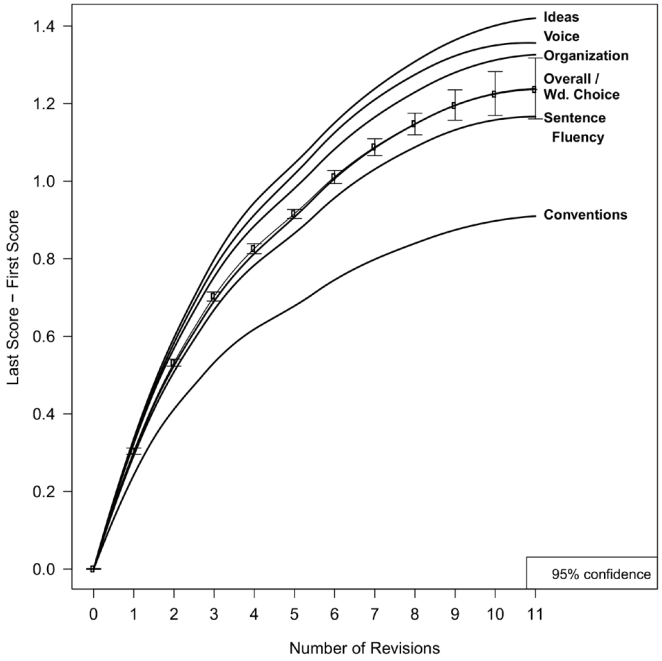

Figure 17.2. Change in writing scores for multiple writing traits across revisions.

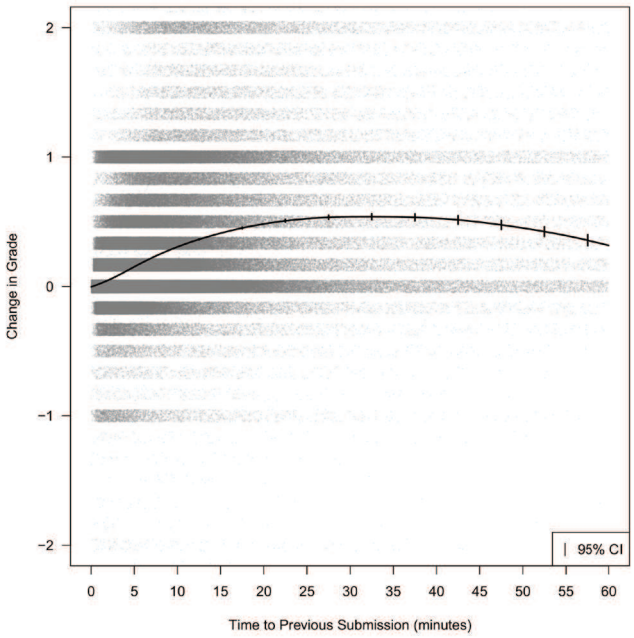

Figure 17.3. Change in grade over revisions based on the time to revise.

tive change (essays receiving a lower score than the previous version) occurs with revisions of less than five minutes. The results suggest an optimal range of time to spend revising before requesting additional feedback. These two results indicate how analysis of log data can validate that the write-feedback-revise cycle improves writing skills, as well as illustrates the ability to fine-tune learning by attempting to lead the student into more effective cycles where feedback is requested at appropriate intervals.

Modelling

The underlying structure of the writing process, as it manifests within a formative writing tool, is often best made interpretable through the construction of formal statistical models. With their explicit representations of the complex interplay of revising, receiving writing advice, and composing responses to multiple prompts over time, these models provide estimates and confidence intervals for parameters of critical interest. Grounded by the student log data, these models can account and control for the complex covariance structure implicit within this stream of data with its aspects of repeated measures of performance received on shared prompts embedded in an overall longitudinal model of growth that can span a significant portion of the total time a student receives writing instruction. A carefully constructed model facilitates teasing out student progress with exposure to the tool, allows placing both students and items on scales of skill level and difficulty respectively, and provides estimates on how changing levels of exposure to components of the available feedback impacts writing performance.

The models described here are based on over 840,000 essays written against more than 190 prompts over a 4-year period by approximately 80,000 students, where over 20% of the students were followed for three or more years. The models predict the holistic score for each essay submitted for feedback, which given the explanatory variables signifies the expected score a student would receive on their essay. The explanatory variables allow us to estimate and control for factors such as the student's grade level, the length of the essay, and the difficulty of the prompt.

The writing process is represented within a linear mixed effects model framework (Pinheiro & Bates, 2006), building on the techniques described in Baayen, Davidson, and Bates (2008). Mixed effects models can estimate both the student's "skill level" and the item's "difficulty" by viewing them as being sampled from a population of all potential students and a bank of all potential prompts, estimates computed in addition to the relationships that hold over the entire population. The students and prompts were modelled as random effects drawn from a distribution with a mean of zero and with the standard deviation estimated from the data. The derived variability provides an estimate of student individual differences, while also capturing the variability of item difficulty. Table 1 contains descriptions of the fixed and random effects used in the models.

At each student grade level, the impact of the higher grade is to increase the score, while as content grade level increases (the labelled grade level of a prompt) the expected score decreases. Finally, in controlling

Table 17.1. Description of Fixed and Random Effects

Variable Name	Description
Fixed Effects	
studentGradeLevel:n	Student's grade level as a factor level (coefficient is the difference between grade n and grade 3)
contentGradeLevel	Grade level of prompt (an assigned level)
log10(wordCount)	Log base 10 of word length of essay
attempt	For a given prompt, the revision of this specific essay submission
elapsedTimeDay	A measure of time in days of how long since first W2L use (a measure of age-based growth)
cumW2LTimeDay	Total face-time student has had with W2L by this submission
interaction	Total number of submissions this student has made to W2L
Random Effects	
studentID	Factor levels, one for each student
contentID	Factor levels, one for each prompt

for essay length, a longer essay on average would be expected to receive a higher score. The four measures of exposure to WriteToLearn™ are all statistically significant and positive, indicating its cumulative positive effect.

While the four measures of WriteToLearn™ are related – e.g., as number of attempts on a specific prompt increases, concurrently the total time spent using WriteToLearn™ increases – they capture different aspects of student interaction with the system. The effect sizes seems small; for instance, each additional attempt on a specific prompt increases the expected score by only .018, a number that represents just the increase based on receiving feedback on a single revision of the essay. In fact, it is only through data mining and modelling with large data sets that we can reliably estimate these important small, incremental effects. From a more global perspective, the cumulative impact of attempts and time spent interacting with WriteToLearn™ result in improvements in achievement. This progress is often best benchmarked with external validations such as those observed in improved pass rates on state achievement tests with more intensive use (Mollette & Harmon, 2015).

Modelling to Determine Writing Prompt Difficulty

Many pedagogical considerations arise in assigning a prompt to a student or a class and one often-expressed concern is adjusting the scoring of the prompt to the student's level (see also Deane & Quinlan, 2010, for a related approach to determining prompt difficulty from a writing corpus). Although some prompts require a threshold skill level or specific knowledge or expertise to be addressed, many are applicable for students over a wide grade range. What differs in the assignment is the expectation of the quality or skill evidenced by the final product and its evaluation via a score. Scoring of prompts is based on grade-specific models, so a prompt labelled as appropriate for 10th graders implies that it is both well-suited for the skills and knowledge expected in 10th grade, but also that the automated scoring was calibrated using training-set essays written by 10th graders. In cases where a prompt is appropriate for a range of grade levels, and training sets of students at different grade levels were available, the same prompt may appear at multiple grade levels, where the critical difference is that different scoring models are used to evaluate the student's work at each grade level.

Often teachers prefer finer levels of discrimination among prompts, such as having a measure of the relative difficulty of a set of prompts that fit the grade level of their class. This is exactly the case that the random effect estimates of the prompts can be used to address. As a prompt's labelled grade level increases, the coefficient on fixed effect contentGradeLevel in the model indicates a 0.073 decrease in expected score (harder prompts contribute to lower scores), other variables held constant. Equivalently, controlling for the labelled prompt grade level, the individual prompt random effects indicates how strongly a given prompt differs in difficulty from this mean fixed effect. This allows ordering the prompts within grade levels, providing

empirically derived additional support infrastructure for teachers. Similarly, taking into account the fixed prompt effect allows ordering all of the prompts, which broadens the set of prompts a teacher may be comfortable assigning.

Title	Grade Level	Difficulty
Freedom of Speech	12	0.80
How to Start a Hobby – A/B	6	0.76
Essay About Causes and Effects in History	11	0.72
How to Start a Hobby	5	0.71
How to Start a Hobby	6	0.70
American President	10	0.68
What's Cooking?	6	0.66
Consumer Reporter	12	0.64
What's Cooking? – A/B	6	0.63
Favorite Activity	4	0.63
...		
Effects of Texting on Communication Skills	8	−0.70
Should Recycling be Voluntary or Required?	7	−0.70
How Much Time to Play Computer Games	7	−0.71
An Unusual Event	9	−0.74
An Important Decision	8	−0.75
Interpret a Literary Theme	7	−0.79
A Meaningful Childhood Memory	10	−0.81
A Meaningful Life Lesson	10	−0.82
Dealing with Conflict	10	−0.85
Compare and Contrast Two Literary Characters	10	−1.08

Beyond this practical result, estimates of prompt difficulty controlling for assigned grade level raise a number of interesting research questions. Table 2 presents a subset of the prompts ordered by the estimates of the conditional modes of the random effects (Bates, Maechler, Bolker, & Walker, 2015) shown in the column called difficulty along with columns for the labelled grade level and the prompt title. The impact on score received on an essay is the sum of the grade level times its coefficient from the model plus its difficulty, so the more positive difficulty is, the easier the prompt is relative to other prompts at that grade level; hence, prompts near the bottom of the table, relative to their grade level, are more difficult. We are just in the early stages of trying to form hypotheses to explain this data, such as why the first 10 prompts in the table are so much easier than other prompts at that grade level and why the last 10 are so much harder, as well as why the relatively easiest items seem to be pulled over a broader range of grade levels than the more constrained set of relatively most difficult items.

Considerations in Modelling with Large Data Sets

In designing a model, there are trade-offs between expressiveness and parsimony. With large datasets, often statistical significance is not a sufficient basis to decide on model form; the purpose of the analysis must also be factored into the decision. A strong message from the descriptive plots presented earlier was that of diminishing returns for variables such as number of submissions per essay. This tendency could be described with a polynomial or in a general additive model context. The power of data mining a large data set is that we can make fewer assumptions about the form relationships will take. In this case, we could assume a linear relation between performance and grade, but instead we estimated a separate improvement relative to 3rd grade, as a baseline, and plotted the relationship in Figure 4. Additional research is necessary to better understand the causes of the asymptotic behaviour and the implications for potential improvements to WriteToLearn™.

We see that from the 4th through about the 10th grade, the improvement is approximately linear, but asymptotes out for 11th and 12th grade. This indicates that at least with this set of prompts and their scoring models, WriteToLearn™ has difficulty distinguishing improvement in writing among 10th through 12th graders. Estimating the slope of the linear portion of the curve from grades 4 to 10 yields a gain of 0.048/grade level, which also can be expressed as an expected gain of 0.29 in going from 4th to 10th grade. This is the expected gain, holding the use of WriteToLearn™ constant. Additional research is necessary to better understand the causes of the asymptotic behaviour and potential improvements to WriteToLearn™.

Related to the work described here are finer grained models of action transitions also using mixed effect models (for instance in a tutoring context see Feng, Heffernan, Heffernan, & Mani, 2009) or using Markov methods (e.g., Beal, Mitra, & Cohen, 2007; Jeong et al., 2008) or Bayesian techniques (e.g., Conati et al., 1997). These techniques can be used to better understand student interactions at the action level (such as use of scaffolding facilities) that complement the more course grained analysis described here.

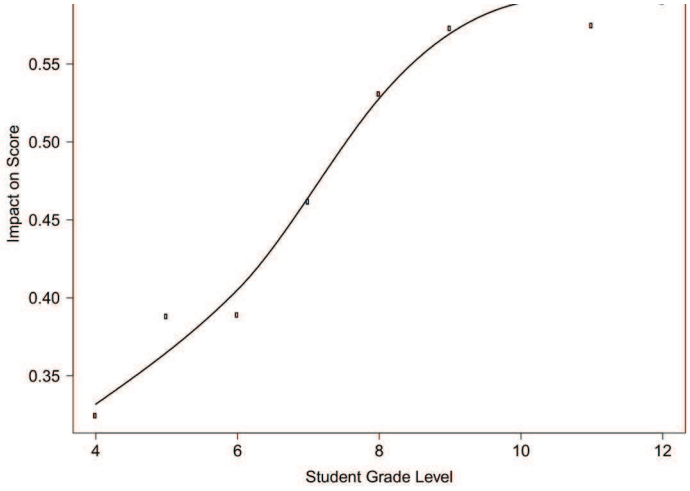

Figure 17.4. Improvement in student score by grade level.

CONCLUSION

Digital education environments can provide an infrastructure to support students with more personalized learning experiences by having students work on more authentic educational tasks while receiving immediate feedback and training specific to their learning needs. Properly instrumented, these environments can also provide a rich source of information about student learning and progress as they interact with the system. Large-scale implementations of formative writing provide rich sets of data for analysis of performance and effects of feedback. By treating the written product as data, applying automated scoring of writing allows monitoring of student learning as students write and revise essays within these implementations. By examining the log of student actions, the amount of time taken, and the changes in the essays, one can track the impact on learning from use of the system.

Developing and maintaining a formative system in a manner to maximize student learning growth requires a range of decisions be made starting from the design and implementation and continuing through the monitoring of its use. Decisions in the design and implementation phase are typically limited to theory and best practices, which are often at a level of granularity that affords a great deal of ambiguity in implementation. However, once a system is deployed, these assumptions can be cast against the actual behaviour of teachers applying the system during their classroom activities and students learning to write. Through data mining, these assumptions can be tested, both to validate the assumptions of the system and to gain greater insight into how students learn.

Writing to Learn and Learning to Write

The resulting analysis validates a key tenet of formative writing: students can improve their writing with revisions based on feedback from the system. A data mining approach to writing permits a fine-grained approach to examining the changes in learning and the effects of feedback on performance. This further permits us to discover, prioritize, and address concerns as they arise and determine which changes are most likely to improve the student experience and their ability to sharpen their writing skills.

The focus of writing assessment has often been put on the product (i.e., the final essay). By performing data mining on student draft submissions and the log of their actions, it is possible to track the process that learners take to create the product. This analysis allows interventions to be performed at strategic points during the process of writing rather than just evaluating the end-product. A wide range of types of analyses can be performed on writing data, including examining the essays, the process to create the essays, as well as the progress of the changes. These approaches can be both descriptive analyses and modelling. While we could not possibly provide a comprehensive discussion on all types of analyses in this chapter, the goal was to illustrate a variety of approaches to show how data mining can provide new ways of thinking about collecting evidence of system and student writing performance and uncover patterns that go beyond those apparent from only observing individual students or classrooms.

REFERENCES

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59, 390–412.

Baikadi, A., Schunn, C., & Ashley, K. (2015). Understanding revision planning in peer-reviewed writing. In O. C. Santos, J. G. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, P. Mitros, J. M. Luna, C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura, & M. Desmarais (Eds.), *Proceedings of the 8th International Conference on Education Data Mining* (EDM2015), 26–29 June 2015, Madrid, Spain (pp. 544 – 548). International Educational Data Mining Society.

Bates, D., Maechler, M., Bolker, B. M., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. ArXiv e-print, *Journal of Statistical Software*, http://arxiv.org/abs/1406.5823.

Beal, C., Mitra, S., & Cohen, P. R. (2007). Modeling learning patterns of students with a tutoring system using Hidden Markov Models. *Proceedings of the 2007 conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work*, 238–245.

Black, P., & William, D. (1998). Assessment and classroom learning. *Assessment in Education*, 5(1), 7–74.

Buckingham-Shum, S. (2013). *Proceedings of the 1st International Workshop on Discourse-Centric Analytics*, workshop held in conjunction with the 3rd International Conference on Learning Analytics and Knowledge (LAK '13), 8–12 April 2013, Leuven, Belgium. New York: ACM.

Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The Criterion Online writing service. *AI Magazine*, 25(3), 27–36.

Calvo, R. A., Aditomo, A., Southavilay, V., & Yacef, K. (2012). The use of text and process mining techniques to study the impact of feedback on students' writing processes. *Proceedings of the 10th International Conference of the Learning Sciences* (ICLS '12) Vol. 1, Full Papers, 2–6 July 2012, Sydney, Australia (pp. 416–423).

Calvo, R. A., O'Rourke, S. T., Jones, J., Yacef, K., & Reimann, P. (2011). Collaborative writing support tools on the cloud. *IEEE Transactions on Learning Technologies*, 4(1), 88–97.

Conati, C., Gertner, A. S., VanLehn, K., & Druzdzel, M. J. (1997). On-line student modeling for coached problem solving using Bayesian networks. *Proceedings of the 6th International User Modeling Conference* (UM97) (pp. 231–242).

Crossley, S. A., McNamara, D. S., Baker, R., Wang, Y., Paquette, L., Barnes, T., & Bergner, Y. (2015). Language to completion: Success in an educational data mining massive open online class. In O. C. Santos, J. G. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, P. Mitros, J. M. Luna, C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura, & M. Desmarais (Eds.), *Proceedings of the 8th International Conference on Education Data Mining* (EDM2015), 26–29 June 2015, Madrid, Spain (pp. 388–392). International Educational Data Mining Society.

Deane, P. (2014). Using writing process and product features to assess writing quality and explore how those features relate to other literacy tasks. *Educational Testing Research Report ETS RR-14-03*. http://dx.doi.org/10.1002/ets2.12002.

Deane, P., & Quinlan, T. (2010). What automated analyses of corpora can tell us about students' writing skills. *Journal of Writing Research*, 2(2), 151–177.

DiCerbo, K. E., & Behrens, J. (2012). Implications of the digital ocean on current and future assessment. In R. Lissitz & H. Jao (Eds.), *Computers and their impact on state assessment: Recent history and predictions for the future*. Charlotte, NC: Information Age.

Feng, M., Heffernan, N. T., Heffernan, C., & Mani, M. (2009). Using mixed-effects modeling to analyze different grain-sized skill models in an intelligent tutoring system. *IEEE Transactions on Learning Technologies*, 2, 79–92.

Foltz, P. W., Gilliam, S., & Kendall, S. (2000). Supporting content-based feedback in online writing evaluation with LSA. *Interactive Learning Environments*, 8(2), 111–129.

Foltz, P. W., & Rosenstein, M. (2013). Tracking student learning in a state-wide implementation of automated writing scoring. *Proceedings of the Neural Information Processing Systems* (NIPS) Workshop on Data Driven

Education. http://lytics.stanford.edu/datadriveneducation/

Foltz, P. W., & Rosenstein, M. (2015). Analysis of a large-scale formative writing assessment system with automated feedback. *Proceedings of the 2nd ACM conference on Learning@Scale* (L@S 2015), 14–18 March 2015, Vancouver, BC, Canada (pp. 339–342). New York: ACM.

Foltz, P. W., & Rosenstein, M. (2016). Visualizing teacher assignment behavior in a statewide implementation of a formative writing system. Cover competition. *Education Measurement: Issues and Practice*, 35(2), 31. http://dx.doi.org/10.1111/emip.12114

Foltz, P. W., Streeter, L. A., Lochbaum, K. E., & Landauer, T. K. (2013). Implementation and applications of the Intelligent Essay Assessor. In M. D. Shermis & J. Burstein, *handbook of automated essay evaluation: Current applications and future directions* (pp. 68–88). New York: Routledge.

Gerbner, G., Holsti, O. R., Krippendorff, K., Paisley, W. J., & Stone, Ph. J. (Eds.) (1969). The analysis of communication content: Development in scientific theories and computer techniques. New York: Wiley.

Graham, S., Harris, K. R., & Hebert, M. (2011). *Informing writing: The benefits of formative assessment*. Carnegie Corporation of New York.

Graham, S., & Hebert, M. (2010). *Writing to read: Evidence for how writing can improve reading*. Carnegie Corporation of New York.

Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology*, 99, 445–476.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112.

Jeong, H., Gupta, A., Roscoe, R., Wagster, J., Biswas, G., & Schwartz, D. (2008). Using Hidden Markov Models to characterize student behaviors in learning-by-teaching environments. In B. Woolf, E. Aïmeur, R. Nkambou, & S. Lajoie (Eds.), *Proceedings of the 9th International Conference on Intelligent Tutoring Systems* (ITS 2008), 23–27 June 2008, Montreal, PQ, Canada (pp. 614–625). Berlin/Heidelberg: Springer.

Krippendorff, K., & Bock, M. A. (2009). *The content analysis reader*. Sage Publications.

Landauer, T. K., Laham, D., & Foltz, P. W. (2001). Automated essay scoring. *IEEE Intelligent Systems*, September/October.

Landauer, T., Lochbaum, K., & Dooley, S. (2009). A new formative assessment technology for reading and writing. *Theory into Practice*, 48(1), 44–52. http://dx.doi.org/10.1080/00405840802577593

Leijten, M., & Van Waes, L. (2013). Keystroke logging in writing research using Inputlog to analyze and visualize writing processes. *Written Communication*, 30(3), 358–392.

Mislevy, R. J., Behrens, J. T., Dicerbo, K. E., & Levy, R. (2012). Design and discovery in educational assessment: Evidence-centered design, psychometrics, and educational data mining. *Journal of Educational Data Mining*, 4(1), 11–48.

Mollette, M., & Harmon, J. (2015). Student-level analysis of Write to Learn effects on state writing test scores. Paper presented at the 2015 annual meeting of the American Educational Research Association. http://www.aera.net/Publications/Online-Paper-Repository/AERA-Online-Paper-Repository

Page, E. B. (1967). The imminence of grading essays by computer. *Phi Delta Kappan*, 47, 238–243.

Parr, J. (2010). A dual purpose data base for research and diagnostic assessment of student writing. *Journal of Writing Research*, 2(2), 129–150.

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert systems with applications*, 41(4), 1432–1462.

Pinheiro, J., & Bates, D. (2006). *Mixed-effects models in S and S-PLUS*. New York: Springer-Verlag.

Reimann, P., Calvo, R., Yacef, K., & Southavilay, V. (2010). Comprehensive computational support for collaborative learning from writing. In S. L. Wong, S. C. Kong, & F.-Y. Yu (Eds.), *Proceedings of the 18th International Conference on Computers in Education* (ICCE 2010), 29 November–3 December, Putrajaya, Malaysia (pp.

129–136). Asia-Pacific Society for Computers in Education.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, 33(1), 135–146.

Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1), 12–27.

Roscoe, R., & McNamara, D. S. (2013). Writing pal: Feasibility of an intelligent writing strategy tutor in the high school classroom. *Journal of Educational Psychology*, 105(4), 1010–1025. http://dx.doi.org/10.1037/a0032340

Shermis, M., & Hamner, B. (2012). Contrasting state-of-the-art automated scoring of essays: Analysis. Paper presented at Annual Meeting of the National Council on Measurement in Education, Vancouver, Canada, April.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, 78(1), 153–189.

Walvoord, B. E., & McCarthy, L. P. (1990). *Thinking and writing in college: A naturalistic study of students in four disciplines*. Urbana, IL: National Council of Teachers of English.

White, B., & Larusson, J. A. (Eds.). (2014). *Learning analytics: From research to practice*. New York: Springer Science+Business Media. doi:10.1007/978-1-4614-3305-7_8.

Whitelock, D., Field, D., Pulman, S., Richardson, J. T. E., & Van Labeke, N. (2013). OpenEssayist: an automated feedback system that supports university students as they write summative essays. *Proceedings of the 1st International Conference on Open Learning: Role, Challenges and Aspirations*. The Arab Open University, Kuwait, 25–27 November 2013.

Whitelock, D., Twiner, A., Richardson, J. T. E., Field, D., & Pulman, S. (2015). OpenEssayist: A supply and demand learning analytics tool for drafting academic essays. *Proceedings of the 5th International Conference on Learning Analytics and Knowledge* (LAK '15), 16–20 March, Poughkeepsie, NY, USA (pp. 208–212). New York: ACM.