Research Article
Coordinated Symposium
NCME 2018 Annual Meeting

Utility-Value Score: A Case Study in System Generalization for Writing Analytics

Beata Beigman Klebanov, *Educational Testing Service*

Stacy Priniski, *University of Wisconsin*

Jill Burstein, *Educational Testing Service*

Binod Gyawali, *Educational Testing Service*

Judith Harackiewicz, *University of Wisconsin*

Dustin Thoman, *San Diego State University*

Abstract

Collection and analysis of students' writing samples on a large scale is a part of the research agenda of the emerging writing analytics community that promises to deliver an unprecedented insight into characteristics of student writing. Yet with a large scale often comes variability of contexts in which the samples were produced—different institutions, different purposes of writing, different author demographics, to name just a few possible dimensions of variation. What are the implications of such variation for the ability of automated methods to create indices/features based on the writing samples that would be valid and meaningful? This paper presents a case study in system generalization. Building on a system developed to assess the expression of *utility value* (a social-psychology-based construct) in essays written by first-year biology students at one postsecondary institution, we vary data parameters and observe system performance. From the point of view of social psychology, all these variants represent the same underlying construct (i.e., utility value), and it is thus very tempting to think that an automatically produced *utility-value score* could provide a meaningful analytic, consistently, on a large collection of essays. However, findings from this research show that there are challenges: Some variations are easier to deal with than others, and some components of the automated system generalize better than others. The findings are then discussed both in the context of the case study and more generally.

Keywords: automated writing evaluation, data variability, first-year STEM, model evaluation, model generalization, STEM motivation. student writing, utility value, writing analytics

1.0 Introduction: Data Variability as a Challenge for Large Scale Analytics

The concept of analytics—meaningful information derived from large-scale data—embodies tension between the increasingly abundant raw data on the one hand and the cost of labor for interpreting the raw data in order to derive analytic indices of interest on the other. In an educational context, for example, it is possible to collect a large number of writing samples that can be used to answer questions such as “What genres of writing are represented in the collection?” or “To what extent do the samples adhere to the writing conventions of English?”. In order to do this, a systematic analysis of the raw data needs to be conducted to create genre labels or scores for adherence to English writing conventions. Asking humans to perform such analyses on a large scale is time-consuming and costly. One way to resolve the cost-vs-scale issue is to combine human and machine “labor” for the large-scale analysis: Humans could provide labels to a small subset of the data, and machine learning would be applied to build an automated system that would extend the labeling to the larger dataset.

An important question that arises in this context is that of the validity of the automatically-produced label. A development plan for an automated system would typically include an evaluation step: A subset of the human-labeled data is deliberately set aside and not used for training the system so that the system can be tested on this new, unseen data. Yet, the accumulation of data for large scale analytics is often dynamic—more data is added in an ongoing, real-time fashion; the question then becomes that of the appropriateness of the automated system that was designed using an early sample for analyzing the incoming data. These new, incoming data might be produced in contexts that are similar but not identical to the earlier sample, in such terms as demographics of the writer population or purpose of writing.

This paper presents a case study in system generalization—specifically, how utility-value scoring models built using one data set generalize to new datasets with varying characteristics. Building on a system originally developed to assess the expression of utility value in essays written by first-year biology students in one postsecondary institution, we vary data parameters, specifically, institution, subject matter, and a detail of the target construct, and observe system performance. From the point of view of social psychology, all these variants represent the same underlying construct (i.e., utility value), and it is thus very tempting to think about an automatically produced *utility-value score* as a meaningful analytic on a large collection of essays. However, we show that this endeavor is not without challenges, as some variations are easier to deal with than others, and some components of the automated system generalize better than others. We discuss the findings both in the context of the case study and more generally.

2.0 Utility-Value Intervention (UVI)

2.1 Background

Keeping students interested in science courses is crucial to retaining them in STEM majors and on track for STEM careers. One way to develop interest in activities is to find meaning and value in those activities (Durik & Harackiewicz, 2007; Hidi & Harackiewicz, 2000). Grounded in expectancy value theory (Eccles & Wigfield, 2002), the utility-value intervention (UVI) aims to promote student motivation and performance by having students reflect on the value of what they are learning. In other words, the UVI seeks to help students focus on the personal relevance and usefulness of the course material, giving them a reason to learn the material (because it is relevant and useful), and therefore increasing their motivation to engage with the material and the likelihood that they will perform well in the course. The intervention typically involves writing assignments, integrated into the curriculum as homework and completed for course credit. In the control conditions, students summarize a topic they've been learning about. In the utility-value conditions, students summarize a topic and explain how the topic is relevant or useful in their own or others' lives. Thus, the assignments all have curricular value, but the utility-value assignments have the added benefit of helping students to find value in what they are learning.

A growing body of evidence suggests that the UVI is effective in science courses. Early tests of the intervention improved grades and interest among high school science students with low expectations of success in their science course (Hulleman & Harackiewicz, 2009) and improved interest among college psychology students with a history of poor performance (Hulleman, Godes, Hendricks, & Harackiewicz, 2010). More recent tests of the intervention have found positive effects on performance in college biology and psychology courses for all students, on average (Canning, Harackiewicz, Priniski, Hecht, Tibbetts, & Hyde, 2018; Harackiewicz, Canning, Tibbetts, Priniski, & Hyde, 2016; Hulleman, Kosovich, Barron, & Daniel, 2017), and especially for students with a history of poor performance (Harackiewicz et al., 2016; Hulleman et al., 2017). The UVI has even helped to close achievement gaps for underrepresented racial/ethnic minority students who were also first-generation college students (Harackiewicz et al., 2016). Finally, initial tests of UVI effects on students' STEM pursuits suggest that the intervention can have positive effects on students' intent to major in STEM fields (Canning et al., 2018; Hulleman et al., 2010).

Given the growing evidence of effectiveness, interest in the UVI is growing among STEM educators and researchers alike. However, the intervention takes a great deal of human labor to implement. First, because the intervention is integrated into the curriculum as course assignments, the assignments need to be evaluated. In a large introductory course, typically with three assignments across the semester, this requires hundreds of hours of grading labor. The grading can be done by professors, teaching assistants, or paid reader-graders, but a certain level of expertise is required. Ideally, graders understand the course material well enough to give feedback on the scientific content of the assignments, and understand the concept of utility value well enough to give students in the utility-value conditions feedback on the utility-value content. Thus, graders typically require training to be able to recognize high-quality utility-value connections and give formative feedback when such connections are lacking. The utility-value feedback is important because science students may not be used to including personal content in their science writing assignments and may require some additional supports to understand whether they are meeting expectations in this regard.

In addition to the grading labor, researchers who implement the UVI typically code the assignments for the quality of the utility-value content, in order to assess implementation fidelity as well as gain insight into the mechanisms driving the effectiveness of utility-value writing. In the cases of the largest field tests of the UVI to date (Canning et al., 2018; Harackiewicz et al., 2016), this coding involved evaluating each assignment on how personal and specific the utility-value content was, on a 0–4 scale. This coding was conducted by a team of 10–15 undergraduate research assistants, working approximately eight hours/week throughout the semester. The training for these research assistants involved group and individual instruction over the course of 2–3 weeks, during which the research assistants learned the coding scheme, coded a training set of assignments, and received individual feedback on the accuracy of their coding until they demonstrated mastery. Each assignment was coded by 2–3 coders, and a master coder then compared coders' scores and resolved any disagreements.

In sum, the UVI as typically implemented requires a significant labor investment, both on the part of the instructional team and the research team. Indeed, the labor investment is likely prohibitive for many instructors who might want to implement the assignment but cannot add so many grading hours to their own (or their instructional staff's) workload. Thus, if the intervention is to be implemented at scale, it will be necessary to develop a less labor-intensive way to evaluate the content of the assignments and give students the necessary feedback to ensure successful and effective implementation of the intervention.

Our previous work with assignments from the Harackiewicz et al. (2016) study suggests that natural language processing (NLP) may provide a useful tool to automate some of the content evaluation process (Beigman Klebanov, Burstein, Harackiewicz, Priniski, & Mulholland, 2017; Beigman Klebanov, Burstein, Harackiewicz, Priniski, & Mulholland, 2016). However, using such tools at scale would require them to be flexible to implementation of the UVI across courses, contexts, and variations of the utility-value assignments. The current paper is the first test of whether previously-developed models and linguistic indicators of utility value can generalize across multiple implementations of the UVI.

The study leverages data collected from UVI studies conducted at California State University-Long Beach, University of Wisconsin-Madison, and at several two-year campuses of the University of Wisconsin Colleges system. We present a series of experiments designed to evaluate generalization of models and linguistic indicators of utility to (a) data from a different institution; (b) data from a different subject (biology vs. psychology); (c) data written in response to a slightly modified utility-value task (personal vs. community utility).

2.2 UVI Tasks and Scoring Rubric

2.2.1 UVI task in institutions A and B. This assignment was administered to introductory biology students at institution A and introductory biology and psychology students at institution B.

> *Assignment: Select a concept or issue that was covered in lecture and formulate a question (see examples at the end of this document). This question should be stated explicitly in your assignment, either as the title or in the first paragraph. Select the relevant information from class notes and the textbook, and write a 500-600 word essay (double-spaced).*
>
> *Write an essay addressing this question and discuss the relevance of the concept or issue to your own life.*[1] *Be sure to include some concrete information that was covered in this unit,* ***explaining why this specific information is relevant to your life or useful for you****. Be sure to explain how the information applies to you personally and give examples.*

Students were given examples of UV connections, relevant to the current course topics. The following were given as examples of the part of an essay that explains personal usefulness:

> *Biology*: “This week we've been talking about osmosis in class and I finally realized why my dad told me once to use honey when I cut myself shaving, since we were out of Neosporin. I thought this was weird until I learned about osmosis and how honey can work as an anti-bacterial ointment, because the sugar to water concentration in the honey is so large that no bacteria can survive. Honey is just one example of how important osmosis is to my life.”
>
> *Psychology*: “Learning about classical conditioning finally made me understand how I developed a fear of bees. I was stung by a bee at a park when I was little. The next year I was stung by a bee again. Both times it really hurt. After that, I became scared of bees. Now I know that the bee was a conditioned stimulus, the sting was an unconditioned stimulus and the fear was the conditioned response. In order to overcome my fear of bees, I need to associate bees with something

[1] Some students were administered a slightly modified version of the task, with this sentence being*: “Write an essay addressing this question and discuss how the information could be useful to you in your own life.”'*

> positive. If I eat candy every time I see one, then the unconditioned stimulus becomes the candy and my happiness becomes the conditioned response."

2.2.2 UV task in institution C data. This writing assignment was administered to students in 1st year biology, chemistry, and physics at institution C.

> *Assignment: Select a concept or issue that was covered in lecture and formulate a question. This question should be stated explicitly in your assignment, either as the title or in the first paragraph. Select the relevant information from class notes and the textbook, and write a 1-2 page essay.*
>
> *Write an essay addressing this question and discuss the relevance of the concept or issue to helping a community to which you belong. Be sure to include some concrete information that was covered in this unit,* ***explaining why this specific information is relevant to, or useful for, helping your community****. Be sure to explain how the information is useful and give examples.*
>
> *Since you will be writing about science from a personal perspective, you should give personal examples and can use personal pronouns (I, we, you, etc.). You do not need to provide citations.*

The following were given as examples of the part of an essay that explains usefulness to community:

> *Biology*: "When I learned about plant diversity in this course, I found out that some plants are more prone to pests than others. Information about plants and their genetics could allow me to know the best ways to control pests. For example, in my textbook I learned about *Bacillus thuringienis*, which is a bacterium that allows for natural resistance against insects in crops. Crops that have resistance against pests could lead to greater food preservation and lower the cost of foods. Now that I have learned more about plant breeding and biotechnology, I think that this information could be used to help people from my own community to have greater access to affordable fresh vegetables and fruits."
>
> *Chemistry*: "When I learned about chemical bonding in this course, I found out that different materials have different properties. Metal materials are held together with metallic bonding, which have good conduction of electricity, malleability, and ductility. Although, metals are very sturdy and can be used for transfer of electricity in a home, they are also relatively expensive. By better understanding ionic bonds, which incorporate metals and nonmetals, we can test physical materials that are sturdy and more cost-effective. Now that I've learned more about chemical bonding, I think that this information could be used to help people from my own community to have more affordable housing."

> *Physics*: “When I learned about momentum in this course, I found out that momentum can be transferred or conserved during motion. I also now know that changes in momentum usually signals a change in velocity. By better understanding how to keep trains at a certain velocity, we can learn how to build more energy efficient trains. More energy efficient public transportation systems, like trains, could lower the cost of transportation. Now that I have learned more about momentum, I think that this information could be used to help people from my own community to have more affordable public transportation.”

2.3 UV Scoring Rubric

The utility-value writing assignments were coded by research assistants for the level of utility value articulated in each essay, on a scale of 0–4, based on how specific and personal the utility-value connection was to the individual. Table 1 shows the rubric.

Table 1

Utility-Value Coding Rubric[2]

UV score	Explanation	Example
0	No UV	This is how muscles work.
1	Non-personal UV	It is important to understand how muscles work.
2	Personal UV, but generic and/or broad, could apply to anyone	I use muscles to move and stay alive.
3	Personal UV with specific connection to a person's life	My friend uses muscles to play soccer.
4	Personal UV with specific connection to a person's life and application, elaboration, or explanation of how that connection matters	Now that I know how muscles build, I know how to optimally train for soccer.

According to Harackiewicz et al. (2016), inter-rater agreement for this coding rubric is high: Two raters provided the exact same score for 91% of the essays. Disagreements were resolved by a master coder following a discussion, if needed. Students were given five days to complete the assignment. Each student contributed three writing samples. This rubric was used at institutions A and B.

For the data collection at institution C, the rubric was modified somewhat to address utility for a specific community rather than a specific individual. Thus, “Knowledge about plant diversity could help with food preservation in my community” would be given a score of 2, since the specified utility could apply to any community. In contrast, the following would be given a

[2] Personal utility may apply to the writer or any other specific “named” individual mentioned in the essay; “personal” utility is not restricted to utility for the self.

score of 4, since the description provides a specific and elaborated connection to the person's community: “My community is facing a severe drought and this causes the price of certain fresh foods to rise. Knowledge about plant diversity could inform my community about which plant-based foods can survive in these conditions and can lead to a more affordable selection of fresh foods to eat.” The inter-annotator agreement was 92% (exact match). Disagreements were resolved by a master coder following a discussion, if needed.

3.0 Data

All writing samples were homework assignments in introductory science courses, completed for course credit. Institution A writing samples were collected from introductory biology courses held between 2012 and 2014. Students were instructed to complete three writing assignments, corresponding to the three main units of the course. Institution B writing samples were collected from introductory biology and psychology courses between 2015 and 2016. In 2015, students were instructed to complete three writing assignments, distributed approximately evenly across the semester. In 2016, students were instructed to complete two writing assignments, in the first 2/3 of the semester. Institution C writing samples were collected from introductory biology, chemistry, and physics courses between 2016 and 2017. Students were instructed to complete three writing assignments, distributed approximately evenly across the semester. The tables below summarize the datasets used in this study, in terms of sizes (Table 2) and distribution of UV scores (Table 3).

Table 2

Description of the Datasets Used in this Study

Institution	Subject	Partition	#Essays
A	Biology	Train	2,766
		Test	329
B	Biology	Test	589
	Psychology	Test	421
C	Biology	Train	427
		Test	187
C	Chemistry	Train	774
		Test	339
C	Physics	Train	671
		Test	293

Note. Train partitions will be used for training automated UV-scoring systems, Test partitions will be used for evaluating the trained models. We use different train-test data combinations for the different experiments reported in this paper.

Table 3

Distribution of Utility Value Scores, Based on Training Data Wherever Applicable

Institution	Subject	UV score				
		0	1	2	3	4
A	Biology	.04	.15	.09	.38	.34
B	Biology	.05	.15	.06	.50	.24
	Psychology	.02	.06	.01	.66	.25
C	Biology	.37	.47	.09	.03	.03
	Chemistry	.42	.44	.08	.03	.03
	Physics	.43	.44	.10	.01	.02

4.0 Baseline Utility-Value Scoring Model

In this section, we provide a high-level description of the features that are used by the model from prior work for predicting the human-assigned utility-value score of a writing sample. This set of features addresses the form and the content of personalized writing; Beigman Klebanov et al. (2017) provide a detailed description. There are 20 features in total.

4.1 Pronouns (5 features)

Grammatical categories that signal self, addressee, or other human reference are typical of UV-rich writing. We thus consider usage of first- and second-person pronouns, possessive determiners (e.g., their) and indefinite pronouns (e.g., anyone).

4.2 General Vocabulary (5 features)

Since expression of UV is likely to refer to everyday concerns and activities, we expect essays rich in UV to be less technical, on average, than essays that only summarize the technical content of a course, and therefore use shorter, more common, and more concrete words, as well as a larger variety of words.

4,3 Argumentative and Narrative Elements (4 features)

While summaries of technical material are likely to be written in an expository, informational style, one might expect the UV elements to be more argumentative, as the writer needs to put forward a claim regarding the relationship between their own or other people’s lives and scientific knowledge, along with necessary qualifications. We therefore defined lists of expressions that could serve to argue and qualify. In addition, in order to connect the science content to the writer's own life, the writer might need to provide a personal mini-narrative – background with details about the events in his or her life that motivate the particular UV statement. A heavy reliance on past test verbs is a hallmark of narrativity, as is use of common action, mental, and desire verbs that could signal sequences of actions and personal stance towards those.

4.4 LIWC (6 features)

We capture specific content and attitude using dictionaries from LIWC (Linguistic Inquiry and Word Count; Pennebaker et al., 2015). In particular, UV statements on biology content often mention the benefit of scientific knowledge for improving understanding and for avoiding unnecessary harm and risk; specific themes often include considerations of health and diet.

5.0 Experiment 1: Institution

In this experiment, we evaluated generalization of the UV Personal Essay scoring models for 1st year biology to a new educational context. Specifically, we trained the models on the A-Biology-Train partition and tested it on the A-Biology-Test partition (same institution) vs. B-Biology-Test partition (different institution). Other than the institution, the setups are comparable, since the relevant studies used the same UV task, exhibited a substantial overlap in the introductory biology topics, and show similar, though by no means identical, UV score distributions, with UV=3 and UV=4 garnering the largest share of essays, followed by UV=1, with fewer essays in UV=0 and UV=2.

Following prior work, we trained a Random Forest regression (Breiman, 2001) using UV features, with Pearson correlation as the objective function. In line with results in Beigman Klebanov et al. (2017), the performance of the model on same-institution dataset was r=0.78. When tested on the B-Biology-Test data, the system performed at r=0.70, exhibiting good generalization.

6.0 Experiment 2: Institution + Subject

In this experiment, we evaluated generalization of the UV Personal Essay scoring models for 1st year biology in one institution to a new subject (1st year psychology) in a different institution. Specifically, we trained the models on the A-Biology-Train partition and tested it on the B-Psychology-Test data. The two data collections used the same UV task. We note that the distributions of UV scores diverge somewhat, as UV=3 dominates with 66% of the data for the Psychology dataset (compared to 38% in the A-Biology-Train dataset), whereas UV=1 that garnered 15% of the data in A-Biology-Train is down to 6% in the B-Psychology dataset.

We used the Random Forest regression model built for experiment 1 above. We observed the correlation of r = 0.57—a substantial correlation, though much reduced from both the original 0.78 and the 0.70 figure from experiment 1.

6.1 Follow-Up: Reduced Feature Set

The different feature sets used in the model can be described as more or less based on the specific content of the essays. Thus, Pronoun and Argumentative/Narrative features capture language that is not specific to the content of discussion. General Vocabulary features capture high-level elements of content, such as the extent to which the writing is academic and/or interpersonal. Finally, LIWC features capture specific content, such as discussion of health or digestion. It stands to reason that when transitioning to a new subject, features that rely on specific content would not generalize as well as features that are more independent of the specific content.

To test this hypothesis, we trained two additional models: (a) No-LIWC, a model with all features apart from LIWC, and (b) No-LIWC-no-General-Vocabulary, a model with only Pronouns and Argumentative/Narrative features. We trained a Random Forest regression on A-Biology-Train data, as before. Consistent with the hypothesis, we found that the No-LIWC model performed at r = 0.60, and the No-LIWC-no-General-Vocabulary model performed at r = 0.62, when evaluated on B-Test-Psychology data.

We may thus conclude that the UV models exhibit a certain extent of cross-subject generalization; among those, the less content-oriented models generalize better.

7.0 Experiment 3: UV Task

We now consider the question whether all or some of the features that were useful in predicting utility-value scores in essays responding to the task administered in institutions A and B would generalize to data collected under a somewhat modified UV task that was used in institution C. The institution C dataset is different in the specifics of the task that is being addressed, as well as in the extent to which essay writers addressed the specific utility required for the high end of the UV score scale. In all subjects in institution C data, 85% or more of the responses received utility-value scores of 0 or 1; namely, the utility value expressed therein is, at most, of the non-personal type (see Table 3).

Given the large discrepancy in the UV score distributions in the A-Biology-train data and all data from institution C, it does not appear promising to test generalization of the UV scoring models from institution A to institution C data. However, it is possible that the features used in the UV model for institution A could be used to discriminate between scores 0 and 1, if the model that would combine them is trained on institution C training data. Our question is therefore—would the features developed for UV scoring models capture the UV distinctions present in a dataset that is collected in a different institution (C vs. A), same but also different subjects (both have biology; institution C also has physics and chemistry), a somewhat modified UV task, and a different empirical distribution of UV scores (scores 3 and 4 dominate in A data; scores 0 and 1 dominate in C data)? To summarize, in this case, we do allow the system to be trained on data that matches the test data in institution (C), subject, and UVI task, but the features extracted from the writing samples were designed for data from a different institution, different subjects, and a slightly different UVI task.

Given the small number of instances with scores 2 and up in institution C datasets, we cast the problem as binary classification to distinguish between score 0 (no UV) and scores 1 and higher (some UV). In order to choose a classifier, we performed cross-validation on each of the three institution C training sets (see Table 2) using a number of classifiers available through the scikit-learn and skll toolkits (Blanchard, Heilman, & Madnani, 2013; Pedregosa et al., 2011). The evaluated classifiers included, among others, Random Forest, Ridge, Linear Support Vectors, AdaBoost, Decision Tree, Gradient Boosting, and Logistic Regression. Two classifiers were selected as top performers for further experiments—Random Forest and Ridge. The classifiers were trained with classification accuracy (proportion correct) as the optimization function.

Table 4 shows the performance of Random Forest and Ridge classifiers trained on the training data for the respective subject and evaluated on test data for the same subject. We observed that while both the selected classifiers consistently outperformed the majority baseline, performance was not very strong, including on biology—a subject for which the features were developed.

Table 4

Binary Classification Accuracies on Test Partitions for Each Subject in Institution C Data

Subject	Majority classifier	Random Forest classifier	Ridge classifier
Biology	.63	.66	.68
Chemistry	.58	.65	.62
Physics	.63	.69	.70

Note. The accuracy of a dummy classifier that always predicts 1 (some UV) is also shown, as Majority; this classifier uses only the distribution of the classes (0 and 1) in the data, without using any of the features of the specific essay.

One possible reason for relatively weak performance (only a few points' improvement over majority baseline) is insufficient training data; the original UV scoring models were trained on thousands of essays, whereas the training set sizes for all institution C data are in the hundreds. While we do not have additional data from the same subject, we could leverage data across subjects, under the hypothesis that expression of utility (even its lowest, non-personal level) would have some linguistic consistency across different subjects. We therefore trained Random Forest and Ridge classifiers on the training data from all subjects—1,872 essays in total—and tested on the combined test data from all the subjects (818 essays). The result was accuracy of 0.67 for each of the two classifiers; that is, we observed similar performance as for the classifiers trained and tested on the same subject. We can tentatively conclude that additional data from same UV task on different subjects does not seem to help.

Another possible reason for the relatively weak performance of the classifiers is that the features were designed for predicting a score on the UV scale of 0–4, not necessarily for identifying essays with the scores of 0 and 1, of which there were not many (see institution A rows in Table 3). Inspecting the published confusion matrix for the UV model used in experiments 1 and 2 (Beigman Klebanov et al., 2017), we observed that of the 31 instances where a human gave 0 (<5% of all essays), the system only identified six correctly—the remaining 25 essays got scores between 1 and 4. Perhaps other features are needed to tell essays with the score of zero (no UV) apart from the rest.

8.0 Discussion and Conclusion

Utility-value intervention aims to promote student motivation and performance by having students reflect on the value of what they are learning. The intervention consists of a series of writing assignments administered after content modules learned in a given subject (such as biology or psychology), about three assignments per semester in the current implementation. A number of field tests showed utility-value intervention to be effective in increasing achievement and motivation, especially for students with a history of poor performance.

Administering a utility-value intervention is, however, quite labor-intensive, in that human evaluators need to be specially trained to recognize expressions of utility value in students' writing and to provide feedback. Such training can make the intervention prohibitively expensive in some educational contexts.

In prior work, we explored the potential of applying natural language processing to students' essays in order to identify indicators of high-quality utility-value writing and create automated models for assessing utility-value in essays. The results of prior work were promising, with correlations of r = 0.78 for the best performing model that leveraged a variety of automatically-computed meaning-based indicators.

Building on our prior work, we further explored the potential of an automated utility-value assessment. We examined the extent to which the model and the features from prior work generalize to data from utility-value interventions administered in a different institution, for different subject matter, and under a modified version of the utility-value task. We found good generalization to data from a new institution (r=0.70 with human-assigned utility-value scores), weaker but still substantial generalization to new subject matter (r=0.62), and relatively weak generalization in the face of a change in utility-value tasks that yielded a substantial difference in the distribution of UV scores from the original data on which the features and models were developed.

Our results suggest that the features identified in prior work do indeed capture linguistic regularities that are associated with expression of utility-value in writing, beyond the specific institution-subject-task setup of the original data. We observed some generalization across all the contexts considered in this paper. That said, our findings clearly paint a rather nuanced picture, in that some features generalize better than others, and some changes of the original setup are more challenging than others, from the point of view of generalization.

We believe that our findings emphasize the need to study the applicability conditions for an automated system if it is going to be tasked with evaluating writing produced in a variety of institutional settings. Thus, it is possible that shifts to a new institution (and, thus, a new student population), a new subject matter course, or a new variant of the original task may result in systematic changes in the textual features that render the original system inapplicable to the new context.

For the wider context of writing analytics, understood here as including systematic analysis of writing samples on a large scale in order to derive meaningful information, our findings suggest caution. In particular, collecting data on a large scale is likely to entail substantial variability in the circumstances where different subsets of the data were produced—different institutional contexts, different points in time, different writer demographics, and different variants of otherwise similar tasks. Such heterogeneity of sources might result in substantial variation in the textual features of the writing samples themselves. This variation, in turn, might pose a problem for automated analysis of the data, as methods developed using earlier, smaller samples might not generalize fully to the larger, and more varied, pool of data; our results provide a case in point. It is therefore important to exercise caution when applying automated methods developed on smaller-scale, substantially homogeneous data to larger, more heterogeneous datasets.

Author Biographies

Beata Beigman Klebanov is a Senior Research Scientist in the Natural Language Processing Group in the Research Division at Educational Testing Service in Princeton, New Jersey.

Stacy Priniski is a National Academy of Education/Spencer Dissertation Fellow and Ph.D. candidate in social psychology at the University of Wisconsin–Madison.

Jill Burstein is a Director of Research for the Natural Language Processing Group in the Research Division at Educational Testing Service in Princeton, New Jersey.

Binod Gyawali is a Research Engineer in the Natural Language Processing Group in the Research Division at Educational Testing Service in Princeton, New Jersey.

Judith Harackiewicz is the Paul Pintrich Professor of Psychology at the University of Wisconsin–Madison.

Dustin Thoman is an Associate Professor in the Department of Psychology and the Center for Research in Mathematics and Science Education at San Diego State University.

Acknowledgments

Research presented in this paper was supported by ETS. Collection of data used in this paper was supported by grants from the National Science Foundation (DRL-1622991) and the National Institutes of Health (2R01GM102703-05). Stacy Priniski is supported by the National Academy of Education/Spencer Dissertation Fellowship and by a grant from the Institute of Educational Sciences, U.S. Department of Education, to the University of Wisconsin-Madison (R305B150003). Any opinions, findings, conclusions, or recommendations expressed in this paper are our own and do not necessarily reflect the views of the ETS, NSF, NIH, or IES. This work was presented at the 2018 Annual Meeting of the National Council on Measurement in Education (NCME) at the Coordinated Symposium entitled: *What Writing Analytics Can Tell Us About Broader Success Outcomes*.

References

Beigman Klebanov, B., Burstein, J., Harackiewicz, J., Priniski, S., & Mulholland, M. (2016). Enhancing STEM motivation through personal and communal values: NLP for assessment of utility value in student writing. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, San Diego, CA.

Beigman Klebanov, B., Burstein, J., Harackiewicz, J., Priniski, S., & Mulholland, M. (2017). Reflective writing about the utility value of science as a tool for increasing STEM motivation and retention – can AI help scale up? *International Journal of Artificial Intelligence in Education*, *27*(4), 791–818.

Blanchard, D., Heilman, M., & Madnani, N. (2013). *SciKit-Learn Laboratory*. GitHub repository, https://github.com/EducationalTestingService/sklI.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Canning, E., Harackiewicz, J., Priniski, S., Hecht, C., Tibbetts, Y., & Hyde, J. (2018). Improving performance and retention in introductory biology with a utility-value intervention. *Journal of Educational Psychology*, *110*(6), 834–849.

Durik, A. & Harackiewicz, J. (2007). Different strokes for different folks: How personal interest moderates the effects of situational factors on task interest. *Journal of Educational Psychology*, *99*, 597–610.

Eccles, J., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, *53*, 109–132.

Harackiewicz, J., Canning, E., Tibbetts, Y., Priniski, S., & Hyde, J. (2016). Closing achievement gaps with a utility-value intervention: Disentangling race and social class. *Journal of Personality and Social Psychology*, *111*, 745–765.

Hidi, S. & Harackiewicz, J. (2000). Motivating the academically unmotivated: A critical issue for the 21st century. *Review of Educational Research*, *70*, 151–179.

Hulleman, C., Godes, O., Hendricks, B., & Harackiewicz, J. (2010). Enhancing interest and performance with a utility value intervention. *Journal of Educational Psychology*, *102*, 880–895.

Hulleman, C. & Harackiewicz, J. (2009). Promoting interest and performance in high school science classes. *Science*, *326*, 1410–1412.

Hulleman, C., Kosovich, J., Barron, K., & Daniel, D. (2017). Making connections: Replicating and extending the utility value intervention in the classroom. *Journal of Educational Psychology*, *109*(3), 387–404.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pennebaker, J. W., Booth, R. J., Boyd, R. L., & Francis, M. E. (2015). *Linguistic Inquiry and Word Count: LIWC2015*. Austin, TX: Pennebaker Conglomerates (www.LIWC.net).