Essays about service-learning events can be mined for program assessment

Anne T. Gilman (gilman@juniata.edu) & Deborah W. Roney (roneyd@juniata.edu)
Juniata College Department of Psychology & Language in Motion, 1700 Moore Street
Huntingdon, PA 16652 USA

Victoria Rehr (vrehr@cscc.edu)
Columbus State Community College Writing Center, 102 Columbus Hall, 550 East Spring Street
Columbus, OH 43215 USA

Helen H. Hu (huh162@psu.edu)
PSU Department of Educational Psychology, Counseling, & Special Education, 125E CEDAR Building
University Park, PA 16802 USA

Mark P. Peterson (mark.phillip.peterson@gmail.com)
Viterbo Universty Department of Biology and Mathematics, 900 Viterbo Drive
La Crosse, WI 54601 USA

Abstract

Psychological applications of human language technology combined with multidisciplinary approaches to similarity calculations and data visualization offer avenues to broaden the use of students' own words in program assessment. We compared multiple analysis approaches on both simple token counts (word roots and character trigrams) and top-down language indicators from 85 student essays about service-learning events. Bioinformatic distance calculations on word root counts provided useable assessment information on attitude change, showing patterns of word use that match the holistic goals of the assignment. Although these patterns were not found in a subsequent batch of 81 essays, the tools we are providing may facilitate other efforts to detect attitude change in student writing about service-learning events.

Keywords: events; semantic similarity; LIWC; text mining

Discovering attitude changes in writing about events

Writing about emotion-laden topics, including events such as starting college, is associated with gains in school success and even physical health (Pennebaker & Chung, 2011). Instructors routinely assess students' content learning and skill development by examining students' assigned writing about events (Reynolds, Livingston, Willson, & Willson, 2010), although questions remain about interpretating these assessment practices (Baker, O'Neil, & Linn, 1993). Institutions of higher learning are facing increasing pressure to provide documentation of changes in knowledge, skill, and attitude in their students (Ewell, 2009). Measuring changes in cultural attitudes is an assessment area with even less consensus than measuring writing skill development. Some measures rely on a few self-reported rating-scale items (Pascarella, Wolniak, Seifert, Cruce, & Blaich, 2005). Other measures that avoid self-report and claim to indicate implicit attitudes can be gamed (Marini, Rubichi, & Sartori, 2012) and should be interpreted with caution (Gawronski, LeBel, & Peters, 2007). Student writing constitutes a greater volume than most professors can assess for learning goals beyond the grading criteria for each assignment. Although this volume remains tiny in comparison to the language data used in computational linguistics research, could some computational techniques help document attitude changes and avoid the need for additional assessment activities? Student privacy protection and variation among assignment types stand in the way of producing a useable large-scale corpus of this type of writing. In this study, we used students' reflections on service-learning events to explore the potential for computational linguistics to supplement other evalutations of student learning.

How computers detect patterns in texts

Enormous growth (Lyman & Varian, 2003) in available, machine-readable text data continues to foster the development of tools for detecting patterns in a much greater volume of writing than a single instructor could feasibly read (Bender & Good, 2010), These tools count written words, phrases, and/or pieces of words in a given text and apply different calcuations using those counts to compare that text to others. For example, "happy birthday to you" is more similar by word count to "you are too happy" than to "many happy returns". A knowledge-added approach might have the phrases "happy birthday" and "happy returns" listed in the same category and thus identify those as more similar to each other than to "too happy". With longer and more numerous texts, the calculations used to measure text similarity–usually expressed as a *distance*—become more complex. The calculations that group documents by topic draw on different text characteristics than those that can detect an author's emotional outlook (Pennebaker, Mehl, & Niederhoffer, 2003), mental health status (Resnik, Garron, & Resnik, 2013), or personal identity (Koppel & Schler, 2003). In this study, we compared several similarity calculations, drawing from bioinformatics as well as computational linguistics, to discover common patterns in essays about service-learning events.

Contrasting computational techniques

A key challenge in finding similar patterns in a group of natural language texts is the sparseness of the distribution of individual words. Among one hundred essays about a service trip

abroad, these terms might each occur three times: *undoubtedly*, *counterrevolutionary*, *twin*. How should similarity calculations treat these terms compared to extremely common ones such as *and* or *the*?

Some approaches divide words into short sequences of letters such as the three-letter trigrams used as part of our calculation; others rely on a dictionary to remove endings and count word roots, so that *revolutionary* and *revolutionaries* count as the same term. Individual words, letter sequences, and word roots are called tokens when used this way. Other approaches add even more knowledge by grouping terms by their part of speech (e.g. noun, verb) and even their emotional connotations (e.g. *glad*, *happy*). Since character trigrams have been employed to allow for errors from faulty optical character recognition in scanned documents (Faber, Hochberg, Kelly, Thomas, & White, 1994), they may aid the processing of documents authored by native and non-native speakers. On the other hand, completely bottom-up approaches often fall short of the mark in natural language processing (Chang & Su, 2004), so we also used top-down semantic and rough part-of-speech information via the Linguistic Inquiry and Word Count (LIWC) tool (Francis & Pennebaker, 1993; Pennebaker et al., 2003), as a comparison facilitated by bioinformatic distance measures.

Applying bioinformatic analysis to student writing

Biology, like computational linguistics, is seeing enormous growth in new tools to find patterns in the rapidly-increasing volumes of data, for example (Buonaccorsi et al., 2014; Sboner, Mu, Greenbaum, Auerbach, & Gerstein, 2011). The focus of bioinformatics in particular on deep analysis of often small numbers of unique samples, makes it a promising source of tools and analogies for processing natural language texts. Many bioinformatic tools converge on the analysis of count data, such as the number of copies of a gene or the number of a bacterial species in a sample. While computational linguistics methods are commonly applied to biomedical research (Ananiadou & McNaught, 2006), applying bioinformatic techniques to language data is still in its infancy. Applying these tools to natural language data stands to increase the reach of natural language processing innovations for text-rich social and behavioral sciences.

Summary

In this paper we present ongoing work to evaluate the utility of low-knowledge and expert-informed computational linguistics tools with and without the application of bioinformatic models for educational program assessment. In addition to new findings, we have developed a small R package to ease the entry of researchers to automated textual analysis. The `wordcountWrapper` package is available at `https://bitbucket.org/petersmp/wordcountwrapper` including source code and a description of the included tools.

Method

Language in Motion (LIM)

As a service-learning program, Language in Motion (LIM) brings college students with extensive knowledge of foreign cultures into rural K-12 classes. For more than a decade, the LIM program has formally documented the professional development gains of participating K-12 teachers and their students. Now, program leaders face growing pressure to label and document the effects on the college students involved: international students, multilingual students, and those returning from study abroad programs. Assessments of LIM's impact for participating K-12 sites do not address program goals such as attitude and perspective changes among presenters.

A bottom-up approach to assessment

The source texts for this project are essays written by students to discuss what they learned from one or more educational outreach events. Each text author (all of whom are college students) made several language and culture presentations to K-12 students, typically featuring a language other than English and stories of the presenter's experiences in another country.

Some program goals, such as practice using the featured language, are comparatively easy to measure independently of these essays; measuring the attainment of goals such as gains in cultural awareness is more difficult. The program administrator (DR), a coauthor and our designated expert, looks for evidence in the essays of progress towards goals such as students learning about themselves and gaining a fresh perspective on their culture of origin. Since expert evaluation of written texts does not map transparently onto assessment criteria that outside reviewers can use, several text-mining techniques were compared to see which most closely approached the expert's holistic essay ratings.

Each presenter submits a narrative evaluation of their experience presenting to younger students, and these essays, averaging 821 words per student, offer the opportunity for automated knowledge extraction. For this study, 85 presenter essays from 2008–2012 were analyzed using computational approaches to increase the depth of evaluation analysis and reporting, including identifying effects not yet articulated in the stated mission of LIM and offering avenues to document qualitative observations.

Data processing

Each essay was converted to an anonymous plain-text document. To remove potentially identifiable (i.e. unique) place names from the corpus and standardize the target culutural information, foreign country names and languages were replaced with *ZZTopia* and *ZZTopian*.

The LIM director rated 22 randomly chosen essays for evidence of learning from their program participation. Our analyses compared those essays rated as showing 'Excellent' success (8) in large program goals with all others (14). Unrated

essays, omitted from clustering analyses, provide useful context for our application of the results found here and will allow for follow-up analysis of the value added by our clustering approaches.

Token counts

Character trigrams and word roots were extracted from each student essay using R (R Core Team, 2013). The functions to read, count, and provide context for these analyses are available from the authors as an R package (`https://bitbucket.org/petersmp/wordcountwrapper`). Punctuation and non-standard characters were discarded, and all numbers were all replaced with a single *7*. Root words were identified using the package `SnowballC` (Bouchet-Valat, 2013). Character trigrams were identified and counted, without spaces between words, using the package `tau` (Buchta, Hornik, Feinerer, & Meyer, 2014). Analysis of whole words was similar to word roots, and trigram analysis including spaces was similar to those without (data not shown).

Standard dimensional reduction

Since word root and trigram data is sparsely distributed, dimensional reduction is needed to make comparisons between texts feasible. To do this, we initially used principal component analysis (PCA), linear discriminant analysis (LDA), and k-means clustering to analyze frequency data for trigrams, word roots, and LIWC categories. Token frequencies within each essay were used rather than raw counts to avoid confounds with differences in essay length, as 'Excellent' essays were approximately twice as long as others (1,269 vs 674 mean words, $t(11.425) = 6.78, p<0.0001$). Each method was then compared to expert ratings to determine their value for future program evaluation.

We compared PCA scores (for any component explaining at least five percent of the variance) between ratings, using a t-test, to determine if any of the dimensions separated our groups. To test the value of the LDA discrimination, we performed a jackknife analysis, serially omitting a single rated sample, and then determining if the LDA test accurately placed it after training on the remaining data. K-means clusters were compared to expert ratings using Cohen's kappa statistic (Cohen, 1960) implemented in the R package `fmsb` (Nakazawa, 2014) to determine if rated groups could be reproduced.

Distance measures from biology

Many methods have been developed in ecology to assess the relative abundance of flora and fauna between divergent sites (Leinster & Cobbold, 2012). One of these measures, the Horn similarity index (Horn, 1966), is now being used in bioinformatics due to its ability to accurately and efficiently handle large numbers of items and samples while accounting for variation in total sampling depth (i.e. length of each essay). We calculated Horn similarities for all pairs of samples using the R package `rnaseqWrapper` (Peterson, Malloy, Buonaccorsi, & Marden, 2015).

Horn similarities were then converted to dissimilarity measures and supplied to the non-metric multidimensional scaling function in the R package `vegan` (Oksanen et al., 2013). Briefly, this calculated the two-dimensional representation of the data that most accurately recreates the pair-wise distances to allow plotting of the spatial relationship between samples.

Differential token usage

To determine which tokens were used differentially by 'Excellent' and other essays, we used the R packages `DESeq` (Anders & Huber, 2010) accessed via `rnaseqWrapper` (Peterson et al., 2015). DESeq uses a negative binomial test on count data to determine whether or not two groups (i.e. 'Excellent' and other essays) have differential representation of a given token (i.e. word root, trigram, or LIWC category).

Designed for gene expression data, DESeq accurately models dispersion across samples by comparing similarly used words to increase the accuracy of the test statistic. However, because it is designed for large scale sequencing projects, the test statistic, particularly its false discovery rate corrected q-value, are likely to be overly conservative for this analysis. Therefore, we report uncorrected p-values but note that these results should be interpreted with caution.

Results

Standard dimensional reduction

Conventional dimensional reduction methods failed to provide any insight into the differences between 'excellent' and other essays for word roots and trigrams, but identified some patterns in LIWC categories. Jackknifed LDA prediction of sample ratings was non-significantly worse than chance for word roots and trigrams, and better than chance for LIWC categories (Fisher's exact test, $p = 0.19$, 0.66, and 0.19).

PCA yielded 5 components in word roots, 7 components in trigrams, and 6 components in LIWC categories that each explained at least five percent of variance (for a total of 57%, 55%, and 83% of variance, respectively). None of these differed between 'Excellent' and other ratings (t-test, all $p > 0.05$). For k-means with two clusters, Cohen's kappa was 0.35, 0.05, and 0.21 for word roots, trigrams, and LIWC ($p =$ 0.06, 0.72, and 0.35), respectively.

Distance measures

Horn distance and non-metric dimensional scaling visualization revealed a centralized cluster of essays rated as 'excellent,' with a dispersion of other essays (Figure 1).

Differential token usage

Between 'Excellent' and other essays, 25 trigrams, 5 root words, and 4 LIWC categories significantly differed in usage (Figure 2). Essays rated as excellent used *both* and *look* more frequently than the others, which in turn had higher occurrences of *didn't*, *American*, and words beginning in *cours*. LIWC analysis showed that the top-rated essays used more

(a) Trigrams (b) Word Roots (c) LIWC Categories

Figure 1: Non-metric scaling of pair-wise Horn similarity indices.

negation, in spite of using *didn't* less, and more perception-related terms. Less-successful essays had more longer words and social terms.

Discussion

We developed a toolkit that enables knowledge extraction from collections of free text where traditional clustering approaches fail. This toolkit delineated student accomplishment of stated but hard-to-measure goals in a language outreach program and promises insights for a range of other educational objectives. Of particular note, the addition of two bioinformatic tools, Horn similarities and DESeq, substantially improved these analyses. These tools revealed clustering that were unlikely to be identified by traditional approaches and identified several of the tokens that characterized high-quality essays.

Clustering

Traditional approaches to dimensional reduction and clustering failed to accurately identify groups, but the addition of Horn similarites suggests why this may be the case. PCA, LDA, and k-means clustering attempt to separate groups along a single axis at a time. The distance measures in the non-metric scaling, however, demonstrated that 'Excellent' essays are instead clustering in the middle.

Collecting more essays might reveal separable, dispersed clusters. Cluster dispersion can indicate mention or absence of specific LIM goals, e.g. changing perspective on one's own culture, or simply reflect an extraneous detail such as a specific food name. That is, addition of more sample essays may allow the identification of a single central cluster (the 'Excellent's here) along with multiple edge clusters, each sharing some common set distinguishing it from the high quality central group. The central distribution of 'Excellent' essays, however, impedes discovery of these dispersed clusters.

Furthermore, this centrality may suggest that 'Excellent' essays share a core vocabulary. This could mean that: *a*) Excellents are not including extraneous information, particularly about their assigned country, and/or *b*) Excellents are touching on more of the core concepts from the program. That is, the clusters of other essays may diverge from the central cluster of "Excellent" essays because they are focused on a single culture (and those cultures may cluster) rather than the broader program goals and/or they may focus on a smaller portion of the goals of the program (and cluster by the portion of the program they cover). This type of analysis would, in part, require breaking the anonymity of the analysis and use subject matter expertise to explore those possibilities.

Specific visualization features for large data sets influence what knowledge viewers gain and remember (Ware, Gilman, & Bobrow, 2008). The bioinformatic visualization tools applied here allowed for more meaningful knowledge extraction from a very limited text corpus.

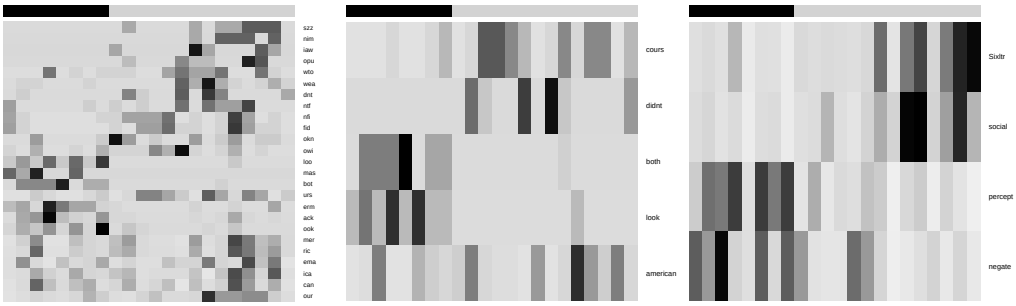
Differential word usage

In addition to this general pattern, we identified specific words and trigrams that are used differently by writers of 'Excellent' versus other essays. In particular, 'Excellent' essays never used the contraction *didn't*, while other ratings used it on average 1.5 times per essay, suggesting that other ratings used less-polished language.

In addition, 'Excellents' used the word root *American* less than other ratings (1.13 vs 3.94 times per essay; similar rates for its component trigrams). This trend continued for focus on the student's country: 'Excellent' essays used the word root *ZZTopia* (an anonymizer for assigned country) less than other ratings (5.61 vs. 8.94 times per essay), though this difference was not statistically significant (p= 0.31). Together, these differences strongly imply that students that are most completely meeting the stated goals of LIM are those that are focusing less on a specific country and more on the cross-cultural goals of the program.

Boundary conditions for LIWC

LIWC analysis did not improve the clustering of essays, though a few categories differed between 'Excellent' and other essays. LIWC can be confounded by departing from the original expressive writing prompt (Hu, Koestler, Stroup, & Gilman, 2013). The present findings confirm that the LIWC tool is distinct from low-knowledge word- and character-based approaches, at the same time demonstrating additional boundary conditions for LIWC.



(a) Trigrams (b) Word Roots (c) LIWC Categories

Figure 2: Heatmap of token usage by rating. For each essay (column), usage of each token (row) significantly differentially used (by DESeq), with usage z-scale transformed by row (darker is higher usage). 'Excellent' essays are to the left in each panel, denoted by the black bars above the heatmap.

No single assessment meets all needs

The educator who provided expert ratings on LIM essays aimed to spend no more than several minutes per essay to arrive at a broad rating of the writer's success in learning from LIM involvement. This rating was specifically not intended to follow from the grade the paper might receive considered only as a piece of writing. Mining assignment text for students' attitude changes should provide different results than expert ratings (Lee, Pincombe, & Welsh, 2005). The tools presented here were applied to a comparison set of 113 essays from British students of anthropology and sociology (Nesi, Gardner, Thompson, & Wickens, 2007), fields where perspective on one's own and others' culture matters. No similar clustering patterns were found according to paper grade or education stage, suggesting that our results are not based on writing quality but on learning from a described event.

Automated essay grading can invite attempts to game such systems (Perelman, 2008). The present study seeks new ways to describe holistic educational outcomes discovered in participants' words—a very different aim than that of grading.

We put 81 LIM essays (23 of which were rated) from 2013-2016 through `wordcountWrapper`. Although the patterns described here were not directly replicated, it is important to note that many of the essays were written by students that went through the program *after* the initial analyses described here were completed. It is possible that changes made in response to insights developed with these tools are impacting the depth and quality of student experiences. In particular, looking at the context of key tokens such as *ZZTopia* revealed that the more recent papers were describing difficult topics about the featured country rather than merely listing topics. The context tools in this package make such comparisons feasible even for hundreds of essays.

Conclusions

Here, we have demonstrated the utility of applying bioinformatic analysis and visualization to text mining for program assessment. Lacking a consensus measure for changes in perspective accompanying service-learning activities, and aware that self-report and implicit measures (including this one) can be gamed, we mined student-authored texts to supplement (not replace) the administrator's assessment data without adding another instrument. Like Paquette, de Carvalho, and Baker (2014), we see this project based on interactions with one expert as a starting point. The R package provides tools to aid novices to NLP analysis the ability to rapidly create substantial count data from a number of formats. The output count tables are intended to be ready to use in any of several analysis and visualization tools, including the ones shown here. In addition, the package has a function to provide context information (i.e., surrounding words) of any token of interest to aid exploratory analyses. Combined, these tools may aid in increasing quantitative reporting to agencies on the effectiveness of difficult-to-assess program objectives across higher education.

Acknowledgments

Portions of this work were supported a summer grant to ATG by the James J. Lakso Endowment for Faculty Excellence and funding to MPP from Penn State University Huck Institues of the Life Sciences and Howard Hughes Medical Institute.

References

Ananiadou, S., & McNaught, J. (2006). *Text mining for biology and biomedicine*. Boston, MA, USA: Artech House.

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*, R106. Retrieved from `http://genomebiology.com/2010/11/10/R106/` doi: 10.1186/gb-2010-11-10-r106

Baker, E. L., O'Neil, H. F., & Linn, R. L. (1993). Policy and validity prospects for performance-based assessment. *American Psychologist*, *48*(12), 1210.

Bender, E., & Good, J. (2010). A grand challenge for linguistics: Scaling up and integrating models. In *Social, behavioral, & economic sciences 2020.* National Science Foundation (USA). (White Paper)

Bouchet-Valat, M. (2013). Snowballc: Snowball stemmers based on the c libstemmer utf-8 library [Computer software manual]. Retrieved from http://CRAN.R-project.org/package=SnowballC (R package version 0.5)

Buchta, C., Hornik, K., Feinerer, I., & Meyer, D. (2014). tau: Text analysis utilities [Computer software manual]. Retrieved from http://CRAN.R-project.org/package=tau (R package version 0.0-16)

Buonaccorsi, V., Peterson, M., Lamendella, R., Newman, J., Trun, N., Tobin, T., ... Roberts, W. (2014). Vision and change through the genome consortium for active teaching using next-generation sequencing (gcat-seek). *CBE-Life Sciences Education*, *13*(1), 1–2.

Chang, J.-S., & Su, K.-Y. (2004). Pitfalls in applying unsupervised learning to NLP. In *IJCNLP-04*. Hainan Island, China.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *37*.

Ewell, P. T. (2009). *Assessment, accountability, and improvement: Revisiting the tensions*. National Institute for Learning Outcomes Assessment (USA). (Occasional Paper No. 1)

Faber, V., Hochberg, J. G., Kelly, P. M., Thomas, T. R., & White, J. M. (1994). Concept extraction: A data-mining technique. *Los Alamos Science*, *22*, 123–149.

Francis, M. E., & Pennebaker, J. W. (1993). *LIWC: Linguistic inquiry and word count* (Tech. Rep.). University of Texas at Austin.

Gawronski, B., LeBel, E. P., & Peters, K. R. (2007). What do implicit measures tell us?: Scrutinizing the validity of three common assumptions. *Perspectives on Psychological Science*, *2*(2), 181-193. doi: 10.1111/j.1745-6916.2007.00036.x

Horn, H. S. (1966). Measurement of 'overlap' in comparative ecological studies. *American Naturalist*, *100*, 419–424.

Hu, H., Koestler, A., Stroup, S., & Gilman, A. (2013). Comparing text characteristics of expressive and values writing. In *Annual meeting of the Eastern Psychological Association.* New York, USA.

Koppel, M., & Schler, J. (2003). Exploiting stylistic idiosyncrasies for authorship attribution. In *Proceedings of ijcai'03 workshop on computational approaches to style analysis and synthesis* (Vol. 69, p. 72).

Lee, M. D., Pincombe, B., & Welsh, M. (2005). A comparison of machine measures of text document similarity with human judgments. In *27th annual meeting of the cognitive science society (cogsci2005)* (pp. 1254–1259).

Leinster, T., & Cobbold, C. A. (2012). Measuring diversity: the importance of species similarity. *Ecology*, *93*(3), 477–489.

Lyman, P., & Varian, H. R. (2003). How much information? 2003. *University of California, Berkeley, California, USA*. Retrieved from http://www.sims.berkeley.edu/how-much-info-2003

Marini, M., Rubichi, S., & Sartori, G. (2012). The role of self-involvement in shifting IAT effects. *Experimental Psychology*, *59*(6), 348–354. doi: 10.1027/1618-3169/a000163

Nakazawa, M. (2014). fmsb: Functions for medical statistics book with some demographic data [Computer software manual]. Retrieved from http://CRAN.R-project.org/package=fmsb (R package version 0.4.3)

Nesi, H., Gardner, S., Thompson, P., & Wickens, P. (2007). The British Academic Written English (BAWE) corpus, developed at the Universities of Warwick.

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., ... Wagner, H. (2013). vegan: Community ecology package [Computer software manual]. Retrieved from http://CRAN.R-project.org/package=vegan (R package version 2.0-10)

Paquette, L., de Carvalho, A. M., & Baker, R. S. (2014). Towards understanding expert coding of student disengagement in online learning. In *Proceedings of the 36th Annual Cognitive Science Conference* (pp. 1126–1131).

Pascarella, E. T., Wolniak, G. C., Seifert, T. A., Cruce, T. M., & Blaich, C. F. (2005). Liberal arts colleges and liberal arts education: New evidence on impacts. *ASHE Higher Education Report*, *31*, 1–168.

Pennebaker, J. W., & Chung, C. K. (2011). Expressive writing: Connections to physical and mental health. *Oxford handbook of health psychology*, 417–437.

Pennebaker, J. W., Mehl, M., & Niederhoffer, K. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, *54*, 547–577.

Perelman, L. (2008). Information illiteracy and mass market writing assessments. *College Composition and Communication*, *60*, 128–141.

Peterson, M., Malloy, J., Buonaccorsi, V., & Marden, J. (2015). Teaching rnaseq at undergraduate institutions: A tutorial and r package from the genome consortium for active teaching. *CourceSource*.

R Core Team. (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/

Resnik, P., Garron, A., & Resnik, R. (2013). Using topic modeling to improve prediction of neuroticism and depression in college students. In *Proceedings of the 2013 conference on empirical methods in natural language processing* (p. 1348-1353). Stroudsburg, PA, USA: Association for Computational Linguistics.

Reynolds, C. R., Livingston, R. B., Willson, V. L., & Willson, V. (2010). *Measurement and assessment in education*. Pearson Education International.

Sboner, A., Mu, X. J., Greenbaum, D., Auerbach, R. K., & Gerstein, M. B. (2011). The real cost of sequencing: higher than you think! *Genome biology*, *12*(8), 125.

Ware, C., Gilman, A. T., & Bobrow, R. J. (2008). Visual thinking with an interactive diagram. In G. Stapleton, J. Howse, & J. Lee (Eds.), *Diagrams* (Vol. 5223, p. 118-126). Springer.

Data collected can be analyzed to assess student learning outcomes for a program. Collective Portfolios: Faculty assemble samples of student work from various classes and use the "collective" to assess specific program learning outcomes. Portfolios can be assessed by using scoring rubrics; expectations should be clarified before portfolios are examined.Â Reflective Essays: generally are brief (five to ten minute) essays on topics related to identified learning outcomes, although they may be longer when assigned as homework. Students are asked to reflect on a selected issue. Content analysis is used to analyze results. Scoring Rubrics: can be used to holistically score any product or performance such as essays, portfolios, recitals, oral exams, research reports, etc. Assessment for learning â€“ the case for formative assessment. This paper provides findings on assessment for learning, drawn from recent analyses undertaken by CERI. It begins with analysis of the formative approach in exemplary practice carried out in secondary schools in eight education systems.Â Note: Information gathered at each level of the system can be used to identify strengths and weaknesses, and to shape strategies for improvement. Source: Authors. Formative assessment â€“ while not a â€œsilver bulletâ€ that can solve all educational challenges â€“ offers a powerful means for meeting goals for high-performance, high-equity of student outcomes, and for providing students with knowledge and skills for lifelong learning.