Developing Component Scores from Natural Language Processing Tools to Assess Human Ratings of Essay Quality

[1]Scott A. Crossley and [2]Danielle S. McNamara

[1]Georgia State University, Applied Linguistics/ESL, 34 Peachtree St. Suite 1200, Atlanta, GA 30303
[2]Arizona State University, Psychology, Learning Sciences Institute, PO Box 8721111, Tempe, AZ 85287
sacrossley@gmail.com, dsmcnamara1@gmail.com

Abstract

This study explores correlations between human ratings of essay quality and component scores based on similar natural language processing indices and weighted through a principal component analysis. The results demonstrate that such component scores show small to large effects with human ratings and thus may be suitable to providing both summative and formative feedback in an automatic writing evaluation systems such as those found in Writing-Pal.

Introduction

Automatically assessing writing quality is an important component of standardized testing (Attali & Burstein, 2006), classroom teaching (Warschauer & Grimes, 2008), and intelligent tutoring systems that focus on writing (e.g., Writing-Pal; McNamara et al., 2012). Traditionally, automatic approaches to scoring writing quality have focused on using individual variables related to text length, lexical sophistication, syntactical complexity, rhetorical elements, and essay structure to examine links with writing quality. These variables have proven to be strong indicators of essay quality. However, one concern is that computational models of essay quality may contain redundant variables (e.g., a model may contain multiple variables of syntactic complexity). Our goal in this study is to investigate the potential for component scores that are calculated using a number of similar indices to assess human ratings of essay quality. Such an approach has proven particularly useful in assessing text readability (Graesser, McNamara, & Kulikowich, 2012), but, as far as we know, has not been extended to computational assessments of essay quality.

To investigate the potential of component scores in automatic scoring systems, we use a number of linguistic indices reported by Coh-Metrix (McNamara, Graesser, McCarthy, & Cai, in press), Linguistic Inquiry and Word Count (LIWC; Pennebaker, Booth, & Francis, 2001) and the Writing Assessment Tool (WAT; McNamara, Crossley, & Roscoe, 2013) as variables in a principal component analysis (PCA). PCA is ideal for our purposes because it uses co-occurrence patterns to transform correlated variables within a set of observations (in this case a corpus of persuasive essays) to linearly uncorrelated variables called principal components. Correlations between the individual indices and the component itself can be used as weights from which to develop overall component scores. We use these component scores to assess associations between the components and human ratings of essay quality for the essays in the corpus. These correlations assess the potential for using component scores to improve or augment automatic essay scoring models.

Automated Writing Evaluation

The purpose of automatic essay scoring (AES) is to provide reliable and accurate scores on essays or on writing features specific to student and teacher interests. Studies show that AES systems correlate with human judgments of essay quality at between .60 and .85. In addition, AES systems report perfect agreement (i.e., exact match of human and computer scores) from 30-60% and adjacent agreement (i.e., within 1 point of the human score) from 85-99% (Attali & Burstein, 2006). There is some evidence, however, that accurate scoring by AES systems may not strongly relate to instructional efficacy. For instance, Shermis, Burstein, and Bliss (2004) suggest that students' use of AES systems results in improvements in writing mechanics but not overall essay quality.

Automated writing evaluation (AWE) systems are similar to AES systems, but provide opportunities for students to practice writing and receive feedback in the classroom in the absence of a teacher. Thus, feedback mechanisms are the major advantage of such systems over AES systems. However, AWE systems are not without fault. For instance, AWE systems have the potential to

overlook infrequent writing problems that, while rare, may be frequent to an individual writer and users may be skeptical about the system's feedback accuracy, negating their practical use (Grimes & Warschauer, 2010). Lastly, AWE systems generally depend on summative feedback (i.e., overall essay score) at the expense of formative feedback, which limits their application (Roscoe, Kugler, Crossley, Weston, & McNamara, 2012).

Writing-Pal (W-Pal)

The context of this study is W-Pal, an interactive, game-based tutoring system developed to provide high school and entering college students with writing strategy instruction and practice. The strategies taught in W-Pal cover three phases of the writing process: prewriting, drafting, and revising. Each of the writing phases is further subdivided into modules that provide more specific information about the strategies. These modules include *Freewriting* and *Planning* (prewriting); *Introduction Building*, *Body Building*, and *Conclusion Building* (drafting); and *Paraphrasing*, *Cohesion Building*, and *Revising* (revising). The individual modules each include a series of lesson videos that provide strategy instruction and examples of how to use these strategies. After watching the lesson videos in each module, students have the opportunity play practice games that target the specific strategies taught in the videos and provide manageable sub-goals.

In addition to game-based strategy practice, W-Pal provides opportunities for students to compose entire essays. After submitting essays to the W-Pal system, students' receive holistic scores for their essays along with automated, formative feedback from the AWE system housed in W-Pal (Crossley, Roscoe, & McNamara, 2013). This system focuses on strategies taught in the W-Pal lessons and practice games. For instance, if the student produces an essay that is too short, the system will give feedback to the user about the use of idea generation techniques such as freewriting. In practice, however, the feedback provided by the W-Pal AWE system to users can be repetitive and overly broad (Roscoe et al., 2012). The repetition of broad feedback is often a product of the specificity of many of the indices included in the W-Pal scoring algorithms. These indices, while predictive of essays quality, can lack utility in providing feedback to users. For instance, the current W-Pal algorithm includes a number of indices related to lexical and syntactic complexity. However, many of the indices are too fine-grained to be practical. For instance, advising users to produce more infrequent words or more infinitives is not feasible because it will not lead to formative feedback. As a result, some of the feedback given to W-Pal users is necessarily general in nature, which may make the feedback less likely to be used in essay revision. One of the goals of the current study is to explore the utility of component scores. Components combine similar variable into one score, which may prove powerful providing summative and formative feedback on essays

Method

Corpus

Our corpus was comprised of 997 persuasive essays. All essays were written within a 25-minute time constraint. The essays were written by students at four different grade levels (9^{th} grade, 10^{th} grade, 12 grade, and college freshman) and on nine different persuasive prompts. The majority of the essays were typed using a word processing system. A small number were hand written.

Human Scores

At least two expert raters (and up to three expert raters) with at least 2 years of experience teaching freshman composition courses at a large university rated the quality of the essays using a standardized SAT rubric (see Crossley & McNamara, 2011, for more details). The SAT rubric generated a rating with a minimum score of 1 and a maximum of 6. Raters were informed that the distance between each score was equal. The raters were first trained to use the rubric with 20 similar essays taken from another corpus. The final interrater reliability for all essays in the corpus was $r > .70$. The mean score between the raters was used as the final value for the quality of each essay. The essays selected for this study had a scoring range between 1 and 6. The mean score for the essays was 3.03 and the median score was 3. The scores were normally distributed.

Natural Language Processing Tools

We initially selected 211 linguistic indices from three NLP tools: Coh-Metrix, LIWC, and WAT. These tools and the indices they report on are discussed briefly below. We refer the reader to McNamara et al. (2013), McNamara et al. (in press), and Pennebaker et al. (2001) for further information.

Coh-Metrix. Coh-Metrix represents the state of the art in computational tools and is able to measure text difficulty, text structure, and cohesion through the integration of lexicons, pattern classifiers, part-of-speech taggers, syntactic parsers, shallow semantic interpreters, and other components that have been developed in the field of computational linguistics. Coh-Metrix reports on linguistic variables that are primarily related to text difficulty. These variables include indices of causality, cohesion (semantic and lexical overlap, lexical diversity, along with incidence of connectives), part of speech and phrase tags (e.g., nouns, verbs, adjectives), basic text measures (e.g., text, sentence, paragraph length), lexical

sophistication (e.g., word frequency, familiarity, imageability, familiarity, hypernmy), and syntactic complexity (including a replication of the Biber tagger; Biber, 1988).

LIWC. LIWC reports on psychological variables (social, affective, and cognitive), personal variables (leisure, work, religion, home, and achievement) and grammatical variables. Affective variables reported by LIWC relate to emotion words such as sadness, anxiety, and anger while cognitive categories related to certainty, inhibition, and inclusion/exclusion. Grammatical categories include pronouns and words related to past, present, and future.

WAT. WAT computes linguistic features specifically developed to assess student writing. These features include indices related to global cohesion, topic development, n-gram accuracy, lexical sophistication, key word use, and rhetorical features. Cohesion features include LSA measures between paragraph types and LSA measures of relevance. N-gram accuracy features include indices related to n-gram frequency, proportion, and correlation. Rhetorical features include indices such as hedges, conjuncts, amplifiers, and conclusion statements.

Statistical Analysis

Adopting an approach similar to Graesser et al. (2012), we conducted a principle component analysis (PCA) to reduce the number of indices selected from the NLP tools into a smaller number of components comprised of related features. The PCA clustered the indices into groups that co-occurred frequently within the texts allowing for a large number of variables to be reduced into a smaller set of derived variables (i.e., the components).

For inclusion into a component, we set a conservative cut off for the eigenvalues of $\lambda > .35$. This ensured that only strongly related indices would be included in the analysis. For inclusion in the analysis, we first checked that all variables were normally distributed. We then controlled for multicollinearity between variables (defined as $r > .90$) so that selected variables were not measuring the same construct. After conducting the factor analysis, we then used the eigen values in the selected components to create weighted component scores. These component scores were then correlated against the human essay scores to examine potential associations with essay quality.

Results

Assumptions

Of the 211 selected variables, 68 were not normally distributed and were removed from the analysis. The majority of these variables were bag-of-word counts taken from LIWC or the Biber tagger replication found in Coh-Metrix. In both cases, the words that informed the variables were highly infrequent in the essays leading to positively skewed distributions. Of the remaining variables, an additional 20 variables were removed because of strong multicollinearity with another variable. After controlling for both normal distribution and multicollinearity, we were left with 123 variables to include within the PCA.

Principle Component Analysis (PCA)

The PCA reported 32 components with initial eigenvalues over 1. Within these 32 components, there was a clear break in eigenvalue between the seventh and eighth component. These eight components explained approximately 43% of the shared variance in the data (see Table 1). The associated scree plot also indicated a point of inflection on the curve beginning at the eighth component.

Table 1
Variance explained: Initial eigenvalues

Component	Percent of variance	Cumulative variance
1	12.891	12.891
2	6.691	19.582
3	5.044	24.626
4	4.649	29.275
5	4.197	33.472
6	3.652	37.124
7	3.516	40.639
8	2.738	43.378
9	2.532	45.910
10	2.310	48.220

Table 2
Variance explained: Rotated loadings

Component	Percent of variance	Cumulative variance
1	7.194	7.194
2	5.633	12.828
3	5.141	17.969
4	4.595	22.564
5	3.391	25.955
6	3.230	29.185
7	2.679	31.865
8	2.666	34.531
9	2.557	37.088
10	2.477	39.564

Considering the rotated loadings in Table 2, there appeared to be a clear break in eigenvalues between the

sixth and seventh components. These seven components explained 32% of the variance in the essays. Given this collective information, we opted for an 8-component solution when examining the PCA. Each of these components and the indices that inform them along with their weighted scores and correlations with human ratings are discussed below.

Component 1

The indices comprising the first component and their loadings are provided in Table 3. This component seems to capture lexical and nominal simplicity. From a lexical simplicity standpoint, the component included more frequent words, shorter words, more frequent trigrams, more familiar words, more specific words, fewer academic words, and more social words (i.e., common words like *family, people,* and *talk*). From a nominal perspective, the component had lower lexical density, few nouns, fewer nominalizations, and more s-bars (i.e., more sentences that have ellipsed nouns).

A correlation was calculated between the weighted scores for this component for each essay and the human ratings for essay quality. The correlation reported $r(997) = -.359$, $p < .001$, indicative of a moderate negative relation with essay quality.

Table 3
Component 1 indices and loadings

Index	Loading
CELEX frequency (all words)	0.818
Average syllables per word	-0.809
CELEX frequency (content words)	0.748
Proportion spoken trigrams	0.735
Lexical density	-0.709
Word familiarity	0.648
Incidence of all nouns	-0.643
Hypernymy	-0.587
Nominalizations	-0.583
Academic words	-0.514
Incidence of s-bars	0.472
Incidence of singular nouns	-0.461
Minimum CELEX frequency sentence	0.411
Length of noun phrases	-0.408
Social words	0.400

Component 2

The indices comprising the second component and their loadings are provided in Table 4. This component appeared to represent text brevity and common n-gram use. From a brevity perspective, the component loaded shorter texts, texts with fewer word types, and text with fewer sentences. From an n-gram standpoint, the component loaded more frequent trigrams and bigrams for both written and spoken texts.

A correlation between the weighted scores for this component for each essay and the human ratings for essay quality reported $r(997) = -.554$, $p < .001$, indicative of a strong negative relation with essay quality.

Table 4
Component 2 indices and loadings

Index	Loading
Written bigram frequency logarithm	0.880
Number of words	-0.855
Spoken trigram frequency logarithm	0.816
Written bigram frequency	0.803
Type count	-0.802
Number of sentences	-0.793
Written trigram frequency	0.763
Spoken trigram frequency	0.663

Component 3

The indices comprising the third component and their loadings are provided in Table 5. This component encapsulated text cohesion, including higher text givenness, semantic similarity (i.e., between sentences and paragraphs), and referential overlap (i.e., word, stem, and argument overlap), as well as lower lexical diversity.

The correlation between the weighted scores for this component for each essay and the human ratings for essay quality, $r(997) = -.239$, $p < .001$, indicated a weak negative relation with essay quality.

Table 5
Component 3 indices and loadings

Index	Loading
LSA givenness	0.849
LSA sentence to sentence	0.824
LSA paragraph to paragraph	0.806
Content word overlap	0.715
Type token ratio	-0.709
Stem overlap	0.680
Lexical diversity (MTLD)	-0.662
Argument overlap	0.655
Lexical diversity (D)	-0.624

Component 4

The indices comprising the fourth component and their loadings are provided in Table 6. Capturing verbal properties, this component loaded more verb phrases, incidence of infinitives, incidence of simple sentences (which contain a single main verb), and more verb base forms. The component also negatively loaded prepositions, prepositional phrases, and determiners, which are syntactically linked to nouns and not verbs.

The correlation between the weighted scores for this component for each essay and the human ratings for essay quality reported $r(997) = -.314$, $p < .001$, indicating a moderate negative relation with essay quality.

Table 6
Component 4 indices and loadings

Index	Loading
Incidence verb phrases	0.794
Density verb phrases	0.794
Incidence of infinitives	0.713
Incidence of simple sentences	0.664
All verb incidence	0.597
Incidence of preposition phrases	-0.570
Incidence of prepositions	-0.554
Incidence of verb base forms	0.529
Incidence of determiners	-0.442

Component 5

The fifth component (see Table 7) represented word concreteness. The indices that loaded positively into this component included human ratings of concreteness (how concrete a word is), imageability (how imageable a word is), and meaningfulness (the number of associations a word contains).

The correlation between the weighted scores for this component for each essay and the human ratings for essay quality reported $r(997) = .188$, $p < .001$, indicative of a weak positive relation with essay quality.

Table 7
Component 5 indices and loadings

Index	Loading
Word imageability	0.807
Word concreteness	0.751
Word meaningfulness	0.717

Component 6

The sixth component (see Table 8) appeared to encapsulate syntactic simplicity. The component positively loaded syntactically similarity between sentences and paragraphs (at the lexical and phrase level) while negatively loading sentence length and number of words before the main verb (both indicators of syntactic complexity). The correlation between the weighted scores for this component and human ratings for essay quality, $r(997) = .008$, $p > .050$, indicated a negligible relation.

Table 8
Component 6 indices and loadings

Index	Loading
Syntactic similarity across sentences	0.853
Average sentence length	-0.830
Syntactic similarity across paragraphs	0.825
Words before main verb	-0.365

Component 7

The seventh component (see Table 9) seemed to capture future time. The component positively loaded more future words, modal verbs, and discrepancy words, which include modals and words such as *hope, desire,* and *expect*. The component negatively loaded a verb cohesion index likely as a result of a lack of association between verb cohesion and auxiliary verb use. The correlation between the component scores and essay quality, $r(997) = -.217$, $p < .050$, reflected a weak relation with essay quality.

Table 9
Component 7 indices and loadings

Index	Loading
Future words	0.806
Modal verbs	0.789
Discrepancy words	0.623
Verb cohesion	-0.482

Component 8

The eighth component (see Table 10) represented nominal simplicity. The component positively loaded more noun phrases, but negatively loaded adjectives, which can provide complexity to noun phrases (as well as modifying nouns). The correlation between the weighted scores for this component and the human ratings for essay quality, $r(997) = -.163$, $p < .050$, reflected a weak relation with essay quality.

Table 10
Component 8 indices and loadings

Index	Loading
Density noun phrases	-0.756
Incidence of adjectives	0.707
Incidence of noun phrases	-0.643
Incidence of adjectival phrases	0.624

Discussion

This study demonstrates the potential for combining similar linguistic indices into component scores for essays. Our assumption is that that these components will augment or enhance methods of automatic essay scoring. In particular, our interest is in improving our means of providing feedback to writers in the W-Pal system.

The component scores in this study predominantly demonstrated weak to strong relations with human judgments of essay quality, thus providing some indication of their potential in assessing writing proficiency. The components are also relatively salient in terms of their potential for providing formative feedback. This salience results from the components' clear links to interpretable constructs related to essay quality.

We are nonetheless at initial stages of this project. From the perspective of providing summative feedback on essay quality, our future research will incorporate the components within scoring algorithms and evaluate their

combined strength in predicting essay quality. These studies will assess the value of the components versus individual indices, as well as the combination of the components with individual indices. We presume that many of the components that correlated strongly with essay quality can be used to increase the accuracy of automatic essay scoring algorithms.

We are also interested in the potential of using the component scores to augment formative feedback mechanisms in W-Pal. We expect the component scores to facilitate and enhance our ability to provide feedback on specific aspects of essay quality. For instance, the results here indicate that essays that focus on verbal properties at the expense of nominal properties are associated with lower essay quality scores. The combination of indices potentially provides stronger evidence of such a tendency than do individual indices, such as the presence of nouns. As such, this component score might be translated to a recommendation to writers to focus on providing more evidence and examples to support arguments within the essay.

Overall, we see this study as providing exploratory evidence for the strength of component scores in AWE systems. While we presume that such scores can help increase the accuracy of summative and formative feedback, such presumptions certainly need to be tested empirically. Our future work will explore the use of component scores in the development of both summative and formative feedback to essay quality with the objective of improving feedback, as well as improving our understanding of essay writing. We also plan to investigate potential differences between grade levels and prompts.

Acknowledgements

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A080589 to Arizona State University and Grant R305A080589 to the University of Memphis. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

Attali, Y., & Burstein, J. 2006. Automated Essay Scoring with E-rater V.2. *Journal of Technology, Learning, and Assessment*, *4*3.

Biber, D. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Crossley, S. A., & McNamara, D. S. 2011. Text coherence and judgments of essay quality: Models of quality and coherence. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 29th Annual Conference of the Cognitive Science Society.* (pp. 1236-1241). Austin, TX: Cognitive Science Society.

Crossley, S. A., Roscoe, R., & McNamara, D. S. 2013. Using automatic scoring models to detect changes in student writing in an intelligent tutoring system. In McCarthy, P. M. & Youngblood G. M., (Eds.). *Proceedings of the 26th International Florida Artificial Intelligence Research Society (FLAIRS) Conference*. (pp. 208-213). Menlo Park, CA: The AAAI Press.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. 2011. Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher, 40,* 223-234.

Grimes, D., & Warschauer, W. 2010. Utility in a fallible tool: A multi-site case study of automated writing evaluation. *Journal of Technology, Learning, and Assessment*. *8,* 4-43.

McNamara, D. S., Crossley, S. A., & Roscoe, R. 2013. Natural Language Processing in an Intelligent Writing Strategy Tutoring System. *Behavior Research Methods*, *5* (2), 499-515

McNamara, D. S., Graesser, A. C., McCarthy, P., & Cai, Z. In press. *Automated evaluation of text and discourse with Coh-Metrix.* Cambridge: Cambridge University Press.

McNamara, D., Raine, R., Roscoe, R., Crossley, S., Jackson, G., Dai, J., Cai, Z., Renner, A., Brandon, R., Weston, J., Dempsey, K., Carney, D., Sullivan, S., Kim, L., Rus, V., Floyd, R., McCarthy, P., & Graesser, A. 2012. The Writing-Pal: Natural Language Algorithms to Support Intelligent Tutoring on Writing Strategies. In P. McCarthy and C. Boonthum-Denecke Eds., *Applied natural language processing and content analysis: Identification, investigation, and resolution* pp. 298-311. Hershey, P.A.: IGI Global.

Pennebaker, J. W., Booth, R. J., & Francis, M. E. 2007. *LIWC2007: Linguistic inquiry and word count.* Austin, Texas.

Roscoe, R., Kugler, D., Crossley, S., Weston, J., & McNamara, D. S. 2012. Developing pedagogically-guided threshold algorithms for intelligent automated essay feedback. In P. McCarthy & G. Youngblood (Eds.), *Proceedings of the 25th International Florida Artificial Intelligence Research Society Conference* (pp. 466-471). Menlo Park, CA: The AAAI Press.

Shermis, M.D., Burstein, J.C. & Bliss, L. 2004. The impact of automated essay scoring on high stakes writing assessments. *Paper presented at the annual meeting of the National Council on Measurement in Education*, April 2004, San Diego, CA.

Warschauer, M., & Grimes, D. 2008. Automated writing assessment in the classroom. *Pedagogies: An International Journal, 3,* 22-36.