Facilitating Innovations in Writing Research

Rationale

Teaching students to write is a powerful facilitator of their reading comprehension (e.g., Traga Phillippakos & Graham, 2020); science, social studies, and math content acquisition (Graham et al., 2020); and their effectiveness to communicate in their job or career (National Commission on Writing, 2004). Not only is writing a powerful functional facilitator for academic achievement and career attainment, but it is also a social facilitator to give every student a voice in society (826 National, 2021).

Unfortunately, writing is underprioritized in schools and in research. More than 70% of students in grades 4, 8, and 12 are performing below proficient levels of writing achievement, with Black students, Hispanic students, Native American students, multilingual learners, and students with disabilities are all scored systematically lower than their peers (National Center for Education Statistics, 2012). Instead of blaming students, families, or teachers, education researchers need to expand and diversify research on writing and improve dissemination. A group of prominent researchers from diverse approaches to writing agree that writing assessment needs to be rethought and is a key factor to addressing writing development (826 National, 2021).

How is Writing Measured in Schools?

An estimated 85% of teachers in middle and high school identify their writing rubric as guiding their instruction (Applebee & Langer, 2011). Although the specific content of rubrics varies, writing rubrics contain several dimensions that are each awarded several points. For example, one well-researched rubric includes dimensions of purpose, supporting details, introduction, conclusion, organization, cohesion, and language use; and each dimension is scored from 0 to 5 (Troia et al., 2020).

How is the Rubric Evaluated?

The advantages of rubrics are that the content can be adjusted to fit the specific goals of each genre of writing, they serve as a relatively intuitive guide for instruction, and they can represent the specific sociocultural values to a local context (VanSteendem et al., 2012). But how do we discover the appropriate content for a rubric to promote writing development for all students? A larger-grained measurement is needed for research purposes. One of the disadvantages of rubrics is bias in scoring. Many studies have shown that the score assigned on a rubric is not solely based on the elements defined in the rubric. For example, many rubrics seek to measure the quality of the content that students write without penalizing students for handwriting, spelling, and grammar. A meta-analysis was conducted on studies that compared the ratings of essays (rubrics based on content only) that were typewritten and corrected spelling, and grammar compared to the original uncorrected essays. They found that the effect size of the difference in content quality ratings was -0.38 to -1.03 (Graham et al., 2011). This suggests that raters are unconsciously influenced by aspects of writing unrelated to the dimension defined in the rubric.

Another related reason rubrics are open to bias is lower interrater agreement. Even in the highest-stakes assessment, the National Assessment of Educational Progress, the exact agreement on different dimensions ranges from 55% to 80% (NAEP, 2011). In other words, two highly trained raters with expertise in writing are likely to weigh other student features of writing rather than the dimensions in the rubric when scoring. Issues like this have hindered writing research from being published (IES, 2017).

This project proposes an innovative scoring method that has the possibility to facilitate more research by addressing some of these issues.

Paired Comparison Method

In the paired comparison method (Thurstone, 1927), raters choose which essay communicates more effectively from a pair. All essays are paired until a rank order of all the essays is created. However, in a sample of 100 essays, there would need to be 4,950 pairs and finding the essays by hand to compare is cumbersome. However, comparing all pairs is unnecessary. Mercer found stable reliability by comparing 40% of the pairs (Mercer & Canon, 2021; Mercer et al., 2019). The current proposal builds a web application to automate the presentation of pairs to human raters as well as record the results of the pairing (see user interface in Appendix A). Interrater reliability for this method is very high (*r* =.94) and the differentiation between students creates an ideal scale as an outcome for research studies (Mercer & Canon, 2021). This method does not change the level of systematic bias but does facilitate research to identify sources of systematic bias. It is particularly urgent that we increase research in this area because of the proliferation of machine learning that can exacerbate inequities.

The goal of this project is to create an open-source digital scoring tool to facilitate more and more equitable research on writing. The research team will evaluate its utility by examining three research questions on extant datasets.

Research Questions

1. Do paired comparison ratings measure a latent construct of writing as well or better than an analytic quality rubric?
2. Do teachable components of writing relate to paired comparison ratings in the same or different ways as they relate to an analytic quality rubric?

3. Do paired comparisons ratings detect possible areas of bias?

Methods

Web Application Development

The web application (see Appendix A) will replicate the paired comparison process previously conducted by hand by other researchers (Mercer & Cannon, 2021) and verified via personal communication with Dr. Mercer. During development, Dr. Mercer and the research team will provide iterative input. The research team will also document their procedures to train researchers to conduct reliable ratings. The web application will be open source and all supporting materials will be posted on OSF and GitHub sites.

Demonstrating Utility

To demonstrate the utility of the paired comparison application for the two purposes described above, we will conduct secondary data analysis on two extant datasets used in several prior studies but were not coded with the paired comparison method (Sarmiento et al., 2021; Truckenmiller et al., 2020). Dataset 1 contains essays from 175 Grade 5 students from two rural and suburban school districts in Michigan and was previously scored for several 'teachable' components of writing to evaluate each component's influence on writing quality (as measured by an analytic rubric). These components come from a developmental theory of writing (the Direct and Indirect Effects of Writing; Kim & Graham, 2022; Kim & Schatschneider, 2017), which includes transcription fluency, word-level accuracy, sentence-level accuracy, planning, vocabulary, and diversity of sentences. The writing quality of the essays was previously scored using the rubric example above (Troia et al., 2020). Dataset 1 also contains criterion writing achievement measures.

Dataset 2 contains essays from 396 Grades 3 through 5 students from an urban school district in Florida that has a higher number of essays where students employed markers of African American English (AAE). Dataset 2 was scored for word and sentence errors using two sets of rules: one set based on Mainstream American English (MAE) and the other based on AAE.

The research team will code both datasets using the paired comparison web application and merge them into their corresponding datasets for secondary data analysis of Dataset 1 to answer research questions 1 and 2, and Dataset 2 to answer research question 3.

Research Question 1: Do paired comparison ratings measure a latent construct of writing as well or better than an analytic quality rubric?

The previous study created a latent variable score of writing achievement using the Spontaneous Writing Index of the TOWL-4 (factor loading = .80), the Writing score from the MSTEP (factor loading = .68), and the Reading score from the MSTEP (factor loading = .59). Latent factor scores for each student will be used as the dependent variable in the proposed analysis with paired comparison ratings and writing quality rubric (Troia et al., 2020) as independent variables in two separate analyses to determine the variance accounted for by each type of rating. A hierarchical model (PROC MIXED in SAS 9.4) with restricted maximum likelihood will be used to account for nesting within teachers and to obtain unbiased variance estimates. Variance accounted for will be calculated by comparing an unconditional random effects model to a model that adds the independent variable (paired comparison rating). The research team has published studies using these analyses previously; however, if we encounter any issues with the models, we will seek statistical consultation. The previous study showed that the writing quality rubric predicted 36% of the variance in the latent variable of writing achievement (Truckenmiller et al., 2020). We hypothesize that the paired comparison rating detects as much or more variance, which will demonstrate utility for researchers to use as a dependent variable in future studies.

Research Question 2: Do teachable components of writing relate to paired comparison ratings in the same or different ways as they relate to an analytic quality rubric?

We previously found that the instructional targets of planning, transcription fluency, word-level accuracy, and sentence-level accuracy predicted 43% of the variance in performance on a writing quality rubric (Truckenmiller et al., revise & resubmit) and that vocabulary and syntax predicted 19% (Truckenmiller et al., 2021). The relationship of vocabulary and syntax with writing quality was unexpectedly low, especially given that the writing quality metric explicitly points out vocabulary and syntax as key features of the rubric and does not include accuracy. We hypothesized that challenges were due to adequate but relatively low reliability of the quality score and there was not enough variance in the middle of the distribution on the writing quality score (68% of the scores fell between 7 points and 15 points). We hypothesize that the paired comparison method will more reliably discriminate writing performance on a finer grain and allow future researchers to identify the writing instruction targets that lead to more incremental improvements in writing ability. We will use the same analytic process described in Research Question 1.

Research Question 3: Do paired comparisons ratings detect potential areas of bias?

Previous studies suggest that usage of AAE markers in writing disproportionately drives down the rating of writing quality (e.g., Johnson & VanBrackle, 2012). For the proposed study, there are 60 students in Dataset 2 who have markers of AAE in their essay. We will choose 60 other students from the dataset with matching reading comprehension scores to conduct the analysis. We will conduct a logistic regression with group membership (AAE markers or not) regressed on the paired comparison ratings to demonstrate the utility of paired comparisons to detect aspects of writing in which raters may be unconsciously biasing their judgements of writing. More research is urgently needed by scholars with a variety of expertise to identify biasing factors in the perception of writing quality. We hope that making this scoring mechanism widely and freely available will facilitate researchers from a variety of perspectives including developmental theories, sociocultural theories, and social justice approaches. (1789 words).

References

826 National. (2021). *The Truth about Writing Education in America*. Retrieved from https://826national.org/wp-content/uploads/2021/04/Writing-Eduation-in-America_Full-Report.pdf

Applebee, A. N., & Langer, J. A. (2011). A snapshot of writing instruction in middle schools and high schools. *English Journal, 100*, 14-27.

Graham, S., Harris, K. R., & Hebert, M. (2011). Presentation Effects in Scoring Writing. *Focus on Exceptional Children, 44*, 1-12.

Graham, S., Kiuhara, S. A., & MacKay, M. (2020). The effects of writing on learning in science, social studies, and mathematics: A meta-analysis. *Review of Educational Research, 90*(2), 179-226.

Institute for Education Sciences (IES), Technical Working Group. (2017). *Future directions for writing research at the secondary level*. Washington, DC: Author. Retrieved from https://ies.ed.gov/ncer/whatsnew/techworkinggroup/

Johnson, D., & VanBrackle, L. (2012). Linguistic discrimination in writing assessment: How raters react to African American "errors," ESL errors, and standard English errors on a state-mandated writing exam. *Assessing Writing, 17*(1), 35–54. doi:10.1016/j.asw.2011.10.001

Kim, Y.-S. G. & Graham, S. (2022) Expanding the Direct and Indirect Effects Model of Writing (DIEW): Reading–writing relations, and dynamic relations as a function of measurement/dimensions of written composition. *Journal of Educational Psychology114*(2), 215–238. https://doi.org/10.1037/edu0000564

Kim, Y.-S. G., & Schatschneider, C. (2017). Expanding the developmental models of writing: A direct and indirect effects model of developmental writing (DIEW). *Journal of Educational Psychology, 109*, 35–50. doi:10.1037/edu0000129

Mercer, S. H., Keller-Margulis, M. A., Faith, E. L., Reid, E. K., & Ochs, S. (2019). The potential for automated text evaluation to improve the technical adequacy of written expression curriculum-based measurement. *Learning Disability Quarterly, 42*(2), 117-128.

Mercer, S., & Cannon, J. (2021, November 30). Validity of Automated Learning Progress Assessment in English Written Expression for Students with Learning Difficulties. https://doi.org/10.31219/osf.io/yh3z

National Assessment of Educational Progress (NAEP), U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. (2011). Writing Assessment. Available at www.nationsreportcard.gov/writing_2011/

National Center for Education Statistics. (2012). The nation's report card: Writing 2011 (NCES 2012–470). Washington, DC: Institute of Education Sciences, U.S. Department of Education.

National Commission on Writing. (2004). *Writing: A Ticket to work or a ticket out: A survey of business leaders*. Available at www.collegeboard.com

Sarmiento, C. M., Truckenmiller, A. J., & Cho, E. (July, 2021). Middle school students' use of academic language in narrative and informational writing. In H. Gerde (Chair), *Writing assessment: Novel approaches across ages and contexts*. Symposium presented to the Society for the Scientific Study of Reading, Virtual Meeting

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*, 273–286. doi:10.1037/h0070288

Traga Philippakos, Z. A., & Graham, S. (2020). Teaching Writing to Improve Reading Skills. Research Advisory. International Literacy Association.

Troia, G. A., Brehmer, J., Glause, K., Reichmuth, H., & Lawrence, F. R. (2020). Direct and indirect effects of literacy skills and writing fluency on writing quality across three genres. *Education Sciences, 10*, 297. doi:10.3390/educsci10110297

Truckenmiller, A. J., Cho, E., & Troia, G. (revise and resubmit). Expanding assessment practices to instructionally relevant writing components in middle school. *Journal of School Psychology*

Truckenmiller, A. J., McKindles, J. V., Petscher, Y., Eckert, T. L., & Tock, J. L. (2020). Expanding curriculum-based measurement in written expression for middle school. *Journal of Special Education, 54*, 133-145. doi:10.1177/0022466919887

Truckenmiller, A. J., Shen, M., & Sweet, L. E. (2021). The role of vocabulary and syntax in informational written composition in middle school. *Reading and Writing, 34*, 911-943. doi: 10.1007/s11145-020-10099-1

Van Steendam, E. Tillema, M. Rijlaarsdam, G. & Van den Bergh, H. (2012) *Measuring writing: Recent insights into theory, methodology and practices*. Leiden, the Netherlands: Brill.