National Research University Higher School of Economics
Journal of Language & Education Volume 6, Issue 2, 2020

Roscoe, R., Wilson, J., Patchan, M., Chen, D., & Johnson, A. (2020). Modeling Student Evaluations of Writing and Authors as a Function of Writing Errors. *Journal of Language and Education, 6*(2), 147-164. https://doi.org/10.17323/jle.2020.10316

Modeling Student Evaluations of Writing and Authors as a Function of Writing Errors

Rod Roscoe[1], Joshua Wilson[2], Melissa Patchan[3], Dandan Chen[4], Adam Johnson[1]

[1]Arizona State University

[2]University of Delaware

[3]West Virginia University

[4]American Board of Anesthesiology

Correspondence concerning this article should be addressed to Rod Roscoe, Arizona State University, 7271 E. Sonoran Arroyo Mall, Santa Catalina Hall, Ste. 150, Human Systems Engineering, The Polytechnic School, Ira A. Fulton Schools of Engineering, Mesa, AZ 85212, United States of America.
E-mail: rod.roscoe@asu.edu

Writers are often judged by their audience, and these evaluations can encompass both the text and the authors. This study built upon prior research on writing evaluation and error perceptions to examine how interconnected or separable are these judgments. Using a within-subjects design, college students evaluated four essays demonstrating no errors, lower-level errors, higher-level errors, or both types. Evaluations included writing quality traits (e.g., conventions, ideas, organization, sentence fluency, and voice) and author characteristics (e.g., creativity, intelligence, generosity, and kindness). Exploratory factor analyses identified latent constructs within these ratings. One construct, *Writing Quality and Skill*, appeared to combine writing traits and authors' intellectual ability (e.g., intelligence and knowledgeability). The second construct, *Author Personality,* seemed to comprise interpersonal author traits (e.g., kindness and loyalty). The two constructs were significantly and positively correlated. These results suggest that students tended to form holistic impressions of writing quality and authors rather than distinct judgments about individual traits. The spillover onto perceptions of authors' personal characteristics may be representative of latent biases. Student raters were also more sensitive to lower-level errors than higher-level errors. Implications for biases and training related to peer assessment are discussed.

Keywords: college students, writing evaluation, factor analysis, rater bias, writing instruction, peer assessment

Introduction

Writing and writing evaluation are complex processes that require the development of substantial knowledge and meta-knowledge about language, text, genre, composition, and communication (e.g., Elton, 2010; Flower, Hayes, Carey, Schriver, & Stratman, 1986; Olinghouse, Graham, & Gillespie, 2014; Panadero & Jonsson, 2013; Reiff & Bawarshi, 2011; Wang & Engelhard, 2019). One specific application of this expertise pertains to detection and assessment of writing errors. There are numerous prescriptions, genre conventions, or other 'rules of writing' to consider (Devitt, 2004; Hacker & Sommers, 2016; Hyland, 2007), such as rules for spelling, grammar, and punctuation. Similarly, writing genres might specify criteria for evidence and logical reasoning (e.g., argumentative writing) or characterization and plotting (e.g., narrative writing). Moreover, there are many ways to write well (Crossley, Roscoe, & McNamara, 2014), and variations in style and content interact with audience and background (Magnifico, 2010; McNamara, 2013). Writing evaluators must decide when and whether expectations have been violated—which one might refer to as 'writing errors'—and the complex and subjective nature of writing evaluation means that these decisions could be susceptible to bias or other misleading beliefs. Even experienced raters can be influenced by factors such as race, gender, and class—for example, texts written in African-American Vernacular English may be judged as of lower quality than texts written in Standard American English (Godley & Escher, 2012; Johnson & VanBrackle, 2012).

Numerous studies have observed that texts exhibiting lower-level mechanical errors or higher-level semantic and rhetorical errors are evaluated as lower quality (e.g., Breland & Jones, 1982; Figueredo & Varnhagen, 2005; Jeong, Li, & Pan, 2017; Kreiner, Schnakenberg, Green, Costello, & McClin, 2002; Morin-Lessard & McKelvie, 2019; Vignovic & Thompson, 2010). Because writing is subjective, flexible, and expressive, evaluators sometimes explicitly or implicitly assume that the writing directly reflects the writer (akin to a correspondence bias; see Bauman & Skitka, 2010). When a text exhibits many errors, the authors may be judged as less intelligent, less knowledgeable, less conscientious, less caring, and so on. Indeed, studies have shown that the presence of writing errors impacts perceptions of authors' ability and personality in academic writing (Figueredo & Varnhagen, 2005; Kreiner et al., 2002), professional communication (Vignovic & Thompson, 2010), and informal communication (Boland & Queen, 2016; Cox, Cox, & A. D. Cox, 2017).

These findings inspire a particular concern for how *students* evaluate writing (e.g., peer assessment). As novice writing evaluators (see Attali, 2016; Lim, 2011; Weigle, 1998), students may lack broad or deep knowledge about writing and thus may be ill-prepared to detect certain kinds of errors or evaluate them fairly (e.g., avoiding unwarranted personal inferences about authors). The current study builds upon prior research (Johnson, Wilson, & Roscoe, 2017) to further investigate how writing errors influence students' evaluations of writing quality and author characteristics. Whereas the previous study examined whether college students' ratings varied based on error typology, the current work employs factor analytic methods to better understand the nature of the ratings and their interrelations. Specifically, we consider whether students make multiple distinct evaluations versus fewer holistic judgments, and whether observed latent constructs reflect separate or integrated judgments about text and author. This line of research—both prior and current work—has meaningful implications for peer assessment and how students are trained to evaluate writing.

Students as Evaluators of Writing

Peer assessment is a popular and widely-used intervention for incorporating writing and feedback into diverse learning contexts (Li, Xiong, Hunter, Guo, & Tywoniw, 2019; Panadero & Jonsson, 2013; Topping, 1998, 2009). Peer assessment of writing can be used to directly enable and support instruction in writing courses (Fathi & Khodabakhsh, 2019; Gao, Schunn, & Yu, 2019) and can also support learning within and across other disciplines (Oshner & Fowler, 2004), such as math (Sluijsman, Brand-Gruwel, van Merriënboer, & Bastiaens, 2003), science (Patchan, Schunn, & Clark, 2011, 2018), history (Patchan, Charney, & Schunn, 2009), psychology (Patchan, Hawk, Stevens, & Schunn, 2013; Patchan & Schunn, 2016), and more. Evidence suggests that students learn from writing (Ackerman, 1993; Arnold et al., 2017; Bangert-Drowns, Hurley, & Wilkinson, 2004; Klein, 1999) and from evaluating and giving feedback on others' writing (Cho & MacArthur, 2011; Patchan & Schunn, 2015; see also 'learning by teaching,' Roscoe, 2014). Logistically, recruiting students as peer assessors reduces workload burdens for instructors, thus enabling more writing assignments, writing practice, and iterations of feedback and revising. Finally, research has also demonstrated that students can provide overall reliable, valid, and useful evaluations (Falchikov & Goldfinch, 2000; Gielen, Tops, Dochy, Onghena, & Smeets, 2010; Li et al., 2016; Panadero, Romero, & Strijbos, 2013; Ramon-Casas, Nuño, Pons, & Cunillera, 2019; Schunn, Godley, & DeMartino, 2016). In short, there are numerous reasons to encourage peer assessment of writing in educational settings.

Nonetheless, students are neither expert writers nor evaluators. More expert and experienced evaluators draw from a richer, more nuanced, and more comprehensive understanding of writing. However, as novice writing evaluators and developing writers, college students and adolescents possess incomplete and fragmented writing knowledge and skills that impact their ability to evaluate text. In studies that directly compare student assessors to instructors (i.e., novices versus experienced evaluators), students have provided feedback that was similar to the instructors' feedback yet with a few consistent differences. Students' feedback tends to be shorter, more positive, and focuses less often on high prose or substantive issues (Patchan et al., 2009; 2013; Cho, Schunn, & Charney, 2006; Topping, Smith, Swanson, & Elliot, 2000). For example, Varner, Roscoe, & McNamara (2013) compared high school students' self-assigned scores to their instructor's scores, and used natural language processing tools to reveal textual features associated with such ratings. Students focused on a more limited set of superficial characteristics (e.g., average word length and shallow cohesion) than the teacher (e.g., lexical diversity and sophistication, deep cohesion, elaboration, and organization).

Several studies compared students' feedback to their peers versus their instructor's feedback. In general, students were more positive than the instructors, but with increasing expertise (i.e., lower-ability undergraduates vs. higher-ability undergraduates vs. graduate students vs. postgraduate students) this difference decreased (Patchan et al., 2009; Cho, Schunn, & Charney, 2006; Topping et al., 2000). Expertise also affected the focus of the feedback provided. For example, Patchan et al. (2009) compared peer feedback generated by history undergraduates to their history instructor and a writing instructor. The history instructor primarily noted issues with the history content, whereas the writing instructor focused on solutions to high prose issues. The students usually fell somewhere between the two instructors. Patchan et al. (2011) also compared peer feedback generated by physics undergraduates versus their non-native English-speaking graduate student teaching assistants (TAs). The students provided longer comments and focused more often on high prose than the TAs, and they provided feedback about the physics content just as often as the TAs. Overall, students are able to assess writing and writing errors, but tend to focus on superficial issues or complementary issues compared to teachers (e.g., Topping, et al., 2000).

Prior Study

In a previous study, Johnson et al. (2017, and see Method) focused specifically on college students' evaluations of writing as a function of writing error patterns, and further considered how these evaluations extended to judgments about authors' personal characteristics. That study addressed two primary research questions: how do lower- and higher-level errors influence students' ratings of (1) writing quality and (2) author characteristics?

To answer the above questions, the researchers constructed a set of essays that exhibited 'no errors,' only 'lower-level errors' (e.g., spelling, grammar, and punctuation), only 'higher-level errors' (e.g., ideas, argument, and organization), or both. Participating students then rated four essays that each (a) exhibited a distinct error pattern and (b) appeared to be written by different authors. Ratings included eight writing traits (e.g., conventions, organization, sentence fluency, and voice) and eight author traits (e.g., intelligence, generosity, kindness, and knowledgeability). Researchers analyzed trait ratings as distinct judgments—means, standard deviations, and intercorrelations (see Johnson et al., 2017) were reported for each trait and each error pattern. The authors also reported average 'writing trait' and 'author trait' ratings that aggregated all traits within their respective categories.

Johnson et al. (2017) observed that the presence of writing errors led college students to perceive both writing quality *and* authors more negatively. When essays exhibited errors, students gave significantly lower ratings regarding writing traits (e.g., conventions, organization, and sentence fluency, and also lower ratings on eight author traits (e.g., generosity, kindness, and intelligence. These effects were observed for both lower-level mechanical errors (e.g., spelling and grammar) and higher-level conceptual and rhetorical errors (e.g., missing theses, contradictory arguments, and off-topic examples). Importantly, the effects were stronger for lower-level errors. Some college students did not notice the higher-level errors at all (i.e., gave equivalent ratings to essays with 'higher-level errors only' versus essays with 'no errors'). A key finding, however, was that students indeed made unwarranted judgments about authors based on writing errors—there was no reason to infer that a person was less generous, kind, or loyal due to typos or muddled arguments, and yet students appeared to make such inferences.

One limitation of the prior study is that analyses of either separate or aggregate ratings implied assumptions about how the judgments were (or were not) interconnected. It is possible that students generated distinct evaluations for each trait (i.e., eight writing traits and eight author traits). For instance, when rating 'sentence fluency' and 'organization,' students may have considered these text qualities independently. Likewise, students may have made separate judgments about authors' 'kindness' or 'generosity.' An alternative possibility is that students conceptually combined one or more traits—that is, students' assessments of sentence fluency and organization, or of kindness and generosity, may have been driven by a muddled or blended understanding of these constructs. More importantly, the same unanswered questions apply to whether students evaluated writing quality separately from author characteristics. Perhaps students made only a *single* holistic judgment that a text was 'good' or 'bad,' which then influenced their ratings of *all* individual writing and author traits.

A more technical way to frame these questions is in terms of the latent constructs employed by student raters (i.e., factor analysis). Do aspects of writing quality (e.g., conventions and organization) load on one or more latent factors? And, are those factors separate from author traits (e.g., intelligence and generosity)?

Alternatively, perhaps writing quality and author characteristics load on a single latent factor, implying that they are, in practice, a singular assessment construct. A related issue is how various error patterns influence this interplay between writing and author judgments. Are these evaluations more or less interwoven when a text is relatively free of errors, exhibits only lower-level errors (e.g., spelling), only higher-level errors (e.g., illogical arguments), or both kinds of errors? Prior research has found that different error patterns are not perceived equally by student raters (Johnson et al., 2017), and thus the presence of different errors might plausibly affect students' latent assessment constructs.

- RQ1: Do college students' evaluations of writing quality traits load on a single construct, relatively few constructs, or a variety of distinct constructs? In other words, to what extent do writing assessments represent holistic versus nuanced evaluations?
- RQ2: Do college students' evaluations of author traits load on a single construct, relatively few constructs, or a variety of distinct constructs? In other words, to what extent do author assessments represent holistic versus nuanced evaluations?
- RQ3: Do college students' evaluations of writing quality and author characteristics load on separate latent constructs or overlapping latent constructs? In other words, to what extent are writing and author assessments interconnected, perhaps indicating a risk of bias?
- RQ4: Does the presence or absence of lower-level errors or higher-level errors affect observed latent assessment constructs? In other words, do error patterns influence the manner in which students evaluate writing quality or author characteristics (i.e., as distinct versus holistic judgments)?

With respect to peer assessment of writing, answers to these questions have implications for the extent to which students' evaluations of writing may incorporate interpersonal biases and how training might decouple or address this overlap.

Materials and Methods

The current work entails an extended new analysis of previously collected data. Complete details about data collection (i.e., population, sampling, measures, and materials) is reported in Johnson et al. (2017). However, essential methodological details are reiterated here for clarity.

Participants

Undergraduate students (n = 70) from a large university in the southwestern United States were recruited from Introduction to Psychology courses and compensated via course credit. Participants self-reported a mean age of 20–21 years (M = 20.7, SD = 4.7), with 34.3% identifying as female. Participants identified as African-American (2.9%), Asian (15.7%), Caucasian (42.9%), Hispanic (8.6%), Middle Eastern (22.9%), or Other (7.1%, including multiethnic individuals). The sample was primarily freshmen (54.3%), but also included sophomores, juniors, and seniors. Participants reported a range of academic majors including aviation, business, computing, engineering, life sciences, or other/undeclared.

Research Design and Essay Materials

The study employed a one-way, within-subjects design in which all participants read and rated a total of four essays that each demonstrated a different error pattern: *No Errors*, *Low-Level Errors Only*, *High-Level Errors Only*, or *All Errors*. The essays were constructed by the researchers (see Johnson et al., 2017), but participants were informed that each essay was authored by another student. Specifically, the researchers created essays ostensibly written by four different student authors who expressed unique positions, arguments, and examples. Participants were not given information about the supposed student authors' supposed background (e.g., race or native language) or writing tools (e.g., access to spelling and grammar checking software).

To construct the essay stimuli, the researchers initially drafted four original argument essays in response to a prompt on 'patience' that asked, 'Is it better for people to act quickly and expect quick responses from others rather than to wait patiently for what they want?' These initial essays were revised until all or most mechanical and conceptual errors were removed—subsequently referred to as *No Errors* essays. For *Low-Level Errors Only* essays, every paragraph was modified to include errors in spelling and homophones, capitalization, sentence fragments or run-ons, commas or apostrophes, and verb-noun or tense agreement. However, these essays still contained clear thesis statements, topic sentences, and relevant examples. In contrast, *High-Level Errors Only* essays were mechanically correct but modified to exhibit missing thesis statements and topic sentences, missing evidence, off-topic examples, and contradictory evidence. Finally, *All Errors* essays included both error patterns. Altogether, each of the four error patterns was implemented for each of the four 'student authors,' resulting in 16 total essays. All essays were about 600-650 words in length. To confirm that the constructed essays demonstrated the intended experimental conditions, four expert raters categorized the error pattern of all stimulus essays. Raters exhibited 95.3% accuracy (i.e., experts' categorizations matched the intended patterns) and three of the four raters exhibited perfect accuracy, suggesting that the essay creation process was successful.

Participants were randomly assigned to read and rate four essays such that each error pattern and supposed student author were encountered only once. The sequence of error patterns and authors, along with the error-author pairings, were systematically randomized across participants to control for order effects. Importantly, participants had no knowledge of the experimental manipulation or the intended error pattern of the essays.

Measures

Background Survey

Participants reported their age, gender, race/ethnicity, school year, and academic major at the start of the study.

Writing Quality Ratings

Immediately after reading each essay, participants rated eight writing quality traits. Six traits were selected based on the Six Traits Writing Rubric (Spandel, 2000): *Conventions*, *Ideas and Content*, *Organization*, *Sentence Fluency*, *Voice*, and *Word Choice*. Two additional traits, *Enjoyment* and *Persuasiveness*, sought to elicit evaluations of how pleasurable and convincing the essays were, respectively. Participants were introduced to the traits along with brief, concrete descriptions framed as question prompts (see Table 1). However, formal rubric-referenced training was not provided because our aim was to investigate students' perceptions of writing, author, and errors rather than adherence to a rubric (i.e., a detailed rubric might have heavily influenced the perceptions). Participants rated their agreement with a series of statements (see Table 1) on a scale of 1 ('Very Strongly Disagree') to 10 ('Very Strongly Agree') (one statement per trait).

Author Characteristics Ratings

Participants judged eight student author traits: *Creativity*, *Generosity*, *Hard-working*, *Intelligence*, *Kindness*, *Knowledgeability*, *Loyalty*, and *Thoughtfulness* (see Table 1). Several traits were somewhat more intellectual (e.g., Creativity, Intelligence, Knowledgeability, and Thoughtfulness) and others were more interpersonal (e.g., Generosity, Hard-working, Kindness, and Loyalty). Participants again rated their agreement with a series of statements (see Table 1) on a scale from 1 ('Very Strongly Disagree') to 10 ('Very Strongly Agree').

Procedure

Participants completed the study in a single session (60-90 minutes) that included informed consent and all rating tasks. Ratings of each essay were made immediately after reading that essay. Participants reviewed only one essay at a time.

Data Analysis

Exploratory Factor Analysis

To assess whether raters' latent judgments of writing quality and author characteristics were distinct or overlapping (RQ1, RQ2, and RQ3), we conducted an exploratory factor analysis (EFA) using Mplus v.8.0 software (Muthén & Muthén, 1998-2017). Four separate EFAs were conducted for each of the four essay types: No Errors, Low-level Errors, High-level Errors, and All Errors (RQ4). Each EFA included all eight essay quality and eight

Table 1
Writing and Author Traits, Prompts, and Assessment Statement

Traits	Description Prompt	Assessment Statement
Writing Traits		
Conventions	Does the essay show correct use of spelling, capitalization, punctuation, and grammar?	The essay correctly followed writing conventions (spelling, punctuation, and grammar).
Enjoyable	Is the essay enjoyable or interesting to read?	The essay was enjoyable and interesting to read.
Ideas and Content	Does the essay include a clear main idea? Are ideas supported with relevant details?	The essay contained good ideas and content (main ideas and supporting details).
Organization	Is the essay logically organized? Does the essay include a clear introduction and conclusion?	The essay was organized well (structure, introduction, and conclusion).
Persuasive	Is the essay persuasive and convincing?	The essay was persuasive and convincing.
Sentence Fluency	Does the essay have a smooth flow? Does the essay show effective sentence variety?	The essay demonstrated effective sentence fluency (rhythm, flow, and variety).
Voice	Does the essay convey a clear personality? Does the essay demonstrate awareness of the audience?	The essay demonstrated a clear voice (personality and sense of audience).
Word Choice	Does the essay include carefully chosen wording? Does the essay include vivid images?	The essay used effective word choice (precise and vivid wording).
Author Traits		
Creativity	Is the author a creative and innovative person?	The author is a creative person.
Effort	Is the author a hard-working person?	The author is a hard-working person.
Generosity	Is the author a generous and giving person?	The author is a generous person.
Intelligence	Is the author an intelligent and smart person?	The author is an intelligent person.
Kindness	Is the author a kind and caring person?	The author is a kind person.
Knowledge	Is the author a knowledgeable and well-read person?	The author is a knowledgeable person.
Loyalty	Is the author a loyal and supportive person?	The author is a loyal person.
Thoughtfulness	Is the author a thoughtful and reflective person?	The author is a thoughtful person.

author characteristics variables (i.e., 16 variables per EFA). Maximum likelihood with robust standard errors (MLR) was selected as the estimation method due to the small sample size (n = 70) and because skewness and kurtosis statistics for several essay quality variables indicated statistically significant departures from univariate normality. MLR estimation is robust against violations of normality assumptions, and more appropriate for use in small samples, than the default estimation procedure of maximum likelihood (Byrne, 2013).

Given the small sample size, three methods were implemented to evaluate the adequacy of the sample size for exploratory factor analysis (EFA). For each EFA, we first considered the Kaiser-Mayer-Olkin (KMO) statistic, which reports values ranging from 0.00 to 1.00. Within this range, values between 0.70-0.80 signify ‘good’ sampling adequacy, 0.80-0.90 are ‘very good,’ and values above 0.90 are ‘excellent’ (Field, 2013). Second, we considered the number of variables whose communalities were above 0.60. Variable communality is the proportion (ranging from 0.00 to 1.00) of the variance that a measure shares with other measures. MacCallum, Widaman, Zhang, and Hong (1999) stated that when all communalities are above 0.60, smaller sample sizes (n < 100) may be acceptable. Finally, we also considered the number of factor loadings per factor that were ⩾ .60. Guadagnoli and Velicer (1988) stated that a factor with four or more loadings ⩾ 0.60 is reliable regardless of the sample size (see also Beavers et al., 2013). These sampling appropriateness metrics are reported for each analysis.

Prior to proceeding with EFA, it was also necessary to ensure the presence of sufficient covariation among the observed variables. Descriptive statistics are presented in Table 2, and Tables 3 and 4 present the correlation matrices. With a handful of exceptions, correlations were generally moderate to strong, indicating that EFA was appropriate. Given the moderate to large correlations among the measures within each essay type, an oblique rotation was selected, which allows for correlations among the extracted factors. This approach generates a simple structure while allowing the factors to be correlated. Thus, EFA models were estimated to test between one and six latent factors. An upper limit of six factors was selected because that would result in factors with fewer than three variables, which would likely terminate the estimation procedures.

Table 2

Descriptive Statistics of Observed Ratings of Essay Quality and Author Characteristics

Traits	No Errors		Low-Level Only		High-Level Only		All Errors	
	M	SD	M	SD	M	SD	M	SD
Writing Traits								
1. Conventions	8.69	1.47	4.36	2.61	8.28	1.62	4.32	2.74
2. Enjoyable	7.67	1.97	5.54	2.71	7.59	1.89	5.12	2.83
3. Ideas and Content	7.89	1.57	6.37	2.32	7.20	2.03	5.91	2.66
4. Organization	8.39	1.47	6.50	2.75	7.35	2.12	5.64	2.73
5. Persuasiveness	7.80	1.67	5.58	2.74	7.08	2.19	5.10	2.62
6. Sentence Fluency	8.36	1.35	5.41	2.70	7.41	2.06	5.00	2.68
7. Voice	8.03	1.87	6.39	2.48	7.83	1.84	5.85	2.62
8. Word Choice	7.97	1.59	5.85	2.71	7.69	1.83	5.56	2.55
Author Traits								
9. Creativity	6.91	1.83	5.83	2.32	6.52	1.84	5.47	2.27
10. Generosity	6.67	1.46	5.82	2.17	6.07	1.74	5.44	2.05
11. Hard-working	7.68	1.78	5.76	2.40	7.12	2.06	5.50	2.42
12. Intelligence	7.83	1.64	5.64	2.27	7.20	1.83	5.43	2.45
13. Kindness	6.92	1.73	6.02	2.13	6.68	1.78	6.02	2.13
14. Knowledgeability	7.69	1.92	5.83	2.58	7.13	2.09	5.52	2.62
15. Loyalty	7.08	1.81	5.98	2.07	6.55	1.89	5.72	2.18
16. Thoughtfulness	7.70	1.52	6.32	2.22	6.90	1.95	5.72	2.34

Note. N = 70. Traits 1-8 are ratings of essay quality traits. Traits 9-16 are ratings of author characteristics.

Table 3

Correlations among Ratings for No Errors Essays and Low-Level Error Only Essays

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.
1. Conventions	-	.39	.38	.63	.43	.58	.40	.62	.06	.17	.35	.51	.30	.37	.29	.37
2. Enjoyable	.52	-	.66	.53	.59	.69	.65	.54	.59	.31	.38	.57	.25	.52	.13	.45
3. Ideas and Content	.43	.76	-	.58	.81	.65	.61	.35	.51	.29	.36	.42	.30	.56	.17	.50
4. Organization	.49	.69	.75	-	.55	.58	.42	.58	.27	.27	.25	.45	.29	.36	.20	.47
5. Persuasiveness	.46	.85	.80	.76	-	.62	.57	.43	.48	.25	.38	.53	.22	.53	.18	.47
6. Sentence Fluency	.58	.68	.54	.73	.67	-	.69	.64	.40	.41	.45	.66	.31	.56	.18	.40
7. Voice	.47	.67	.67	.75	.68	.71	-	.46	.18	.31	.33	.45	.39	.34	.19	.35
8. Word Choice	.59	.66	.61	.70	.59	.72	.79	-	.23	.33	.45	.66	.37	.46	.28	.27
9. Creativity	.50	.68	.64	.58	.67	.58	.63	.70	-	.48	.29	.45	.31	.53	.13	.31
10. Generosity	.52	.52	.49	.56	.53	.48	.60	.67	.68	-	.33	.38	.76	.37	.54	.54
11. Hard-working	.56	.64	.58	.58	.63	.52	.69	.72	.69	.71	-	.53	.38	.49	.49	.38
12. Intelligence	.57	.69	.72	.63	.70	.55	.65	.73	.75	.68	.77	-	.39	.65	.23	.39
13. Kindness	.28	.45	.49	.53	.46	.37	.51	.51	.55	.80	.56	.58	-	.41	.53	.47
14.Knowledgeability	.64	.68	.67	.56	.68	.55	.62	.67	.75	.63	.71	.87	.48	-	.31	.50
15. Loyalty	.37	.59	.51	.55	.56	.43	.50	.52	.60	.76	.55	.62	.79	.60	-	.42
16. Thoughtfulness	.36	.62	.63	.68	.71	.56	.65	.62	.63	.66	.59	.73	.62	.68	.68	-

Note. Correlations *above* the diagonal are for ratings of No Errors essays. Correlations *below* the diagonal are for ratings of Low-Level Error Only essays. Correlations ⩾ .23 are statistically significant at $p < .05$. Correlations ⩾ .30 are statistically significant at $p < .01$.

Table 4
Correlations among Ratings for High-Level Error Only Essays and All Errors Essays

	1.	**2.**	**3.**	**4.**	**5.**	**6.**	**7.**	**8.**	**9.**	**10.**	**11.**	**12.**	**13.**	**14.**	**15.**	**16.**
1. Conventions	-	.59	.57	.46	.51	.55	.44	.61	.49	.47	.53	.64	.36	.53	.31	.53
2. Enjoyable	.72	-	.76	.69	.79	.69	.64	.72	.70	.60	.63	.70	.61	.67	.51	.69
3. Ideas and Content	.60	.71	-	.74	.81	.68	.61	.69	.67	.57	.68	.70	.48	.65	.45	.59
4. Organization	.56	.61	.78	-	.70	.64	.55	.60	.55	.43	.65	.68	.58	.53	.42	.52
5. Persuasiveness	.63	.71	.86	.80	-	.75	.67	.71	.57	.51	.59	.73	.43	.64	.37	.59
6. Sentence Fluency	.59	.61	.63	.61	.77	-	.70	.77	.54	.48	.49	.64	.42	.42	.30	.55
7. Voice	.49	.59	.69	.71	.68	.72	-	.69	.48	.36	.46	.60	.30	.44	.15	.52
8. Word Choice	.53	.63	.75	.71	.81	.76	.74	-	.53	.49	.55	.73	.42	.56	.29	.51
9. Creativity	.53	.60	.66	.54	.65	.55	.58	.59	-	.69	.73	.66	.55	.71	.48	.64
10. Generosity	.50	.59	.63	.49	.62	.60	.63	.58	.77	-	.63	.59	.72	.65	.61	.62
11. Hard-working	.55	.58	.70	.67	.67	.59	.69	.64	.53	.69	-	.73	.61	.78	.68	.75
12. Intelligence	.55	.57	.71	.59	.72	.62	.58	.62	.76	.72	.74	-	.62	.69	.41	.69
13. Kindness	.35	.41	.54	.44	.53	.48	.47	.56	.62	.75	.58	.61	-	.60	.59	.71
14. Knowledgeability	.62	.66	.79	.64	.77	.66	.63	.68	.67	.70	.81	.83	.53	-	.59	.70
15. Loyalty	.41	.50	.63	.57	.65	.55	.66	.63	.64	.76	.69	.71	.80	.66	-	.67
16. Thoughtfulness	.41	.53	.74	.59	.70	.59	.62	.59	.60	.71	.77	.72	.65	.81	.76	-

Note. Correlations *above* the diagonal are for ratings of High-Level Error Only essays. Correlations *below* the diagonal are for ratings of All Errors essays. Correlations $\geqslant .23$ are statistically significant at $p < .05$. Correlations $\geqslant .30$ are statistically significant at $p < .01$.

When evaluating the results of the EFA analyses, four metrics were inspected to select the optimal factor solution and determine whether to retain or omit a given factor. First, Kaiser's criterion (1974; see also Ruscio & Roache, 2012) was used to select the appropriate number of factors, which retains factors whose eigenvalues are $\geqslant 1.00$. In addition, scree plots were examined to identify the point of inflection. The number of factors above the point of inflection were retained. Second, parallel analysis (Horn, 1965) was used to retain the number of factors whose eigenvalues are larger than the corresponding eigenvalues from the parallel analysis (see also Ruscio & Roache, 2012). Third, factors were retained if corresponding variables demonstrated appreciable loadings in the range of $\geqslant 0.60$. Extracted factors were dropped if they had no corresponding variables with appreciable loadings. Fourth, factor solutions were selected based on their interpretability. Ultimately, EFA is a statistical method to arrive at theoretical understanding of a phenomenon. Therefore, an *interpretable* EFA solution is preferred when there are competing models and the other selection criteria are unclear.

Finally, three model fit indices were inspected to determine models of latent judgments: chi-square statistics, root mean square error of approximation (RMSEA; Steiger & Lind, 1980), and standardized root mean square residual (SRMR). First, non-significant chi-square values indicate a good fitting model. In the absence of a non-significant chi-square value, models with lower values are considered better fitting than models with higher values. Second, when available, RMSEA is a measure of model fit that accounts for the number of parameters in the model. RMSEA values less than 0.05 indicate good fitting models; values greater than 0.10 indicate poor fitting models (Brown, 2006; Kenny, 2014). Third, SRMR evaluates differences between model-implied correlations and the correlations observed in the data (Brown, 2006). SRMR values less than 0.08 indicate good fit; whereas, 0.00 indicates perfect fit.

Results

Overview

Across all four error patterns, a two-factor solution was generally optimal and the observed latent factors were largely consistent (see RQ4). The first and most dominant factor tended to include all or most of the eight writing quality traits, along with ratings of the student authors' intellectual ability to a lesser extent (e.g., intelligence, knowledge, or creativity). Thus, in answer to RQ1, students generally made holistic judgments about writing quality rather than distinct judgments along one or more writing traits. However, in answer to RQ3, it appears that this construct also incorporated aspects of author evaluations. For this reason, this factor might be labeled W*riting Quality and Skill*—a holistic evaluation of the quality of the essay that also reflects authors' writing ability and mental resources. The second and less dominant factor consistently included interpersonal student author traits perhaps indicative of authors' personality (i.e., generous, kind, or loyal) and sometimes included traits related to effort or conscientiousness. In answer to RQ2, it appeared that students made holistic judgments about authors' interpersonal characteristics rather than nuanced evaluations of individual qualities. Overall, this second factor might be labeled *Author Personality*.

Writing Quality and Skill was consistently and positively correlated with Author Personality. Thus, although they may be separate judgments, they do influence each other. Poor quality writing reflects negatively on the author as a person; writers who seem less kind or generous may inspire readers to be much more critical. With regard to RQ4, consistency across error patterns suggested that the presence of different errors did not dramatically change how students rated writing or authors, although semantic and rhetorical errors may inspire broader judgments of the authors' personality. That is, student raters may view such errors as more 'revealing' about the author than are spelling or grammatical errors. This subtle pattern is discussed further in results for each error pattern.

Writing and Author Ratings of 'No Errors' Essays

For the No Errors essays ratings, metrics of sampling adequacy indicated that it was appropriate to proceed with EFA. The KMO statistic was 0.83 ('very good') and 13 of the 16 variables exhibited communalities greater than or equal to 0.60—only slightly below the criterion suggested by MacCallum et al. (1999).

Analyses considered models from one to six latent factors. All models converged except the six-factor model, which was omitted (see Table 5). After considering all criteria, a two-factor model was selected as the optimal and most parsimonious model. Kaiser's criterion and the scree plot indicated a three-factor model, but the parallel analysis favored a two-factor model. Inspection of the factor loadings also favored a two-factor model. All models yielded a significant chi-square statistic, which indicated possible model misspecification. RMSEA was below the 0.10 threshold only for the five-factor model, and SRMR was acceptable for the four-factor model (0.04) and the five-factor model (0.03).

The pattern matrix for the two-factor solution is presented in Table 6. Nine variables loaded on the first factor with values of 0.60 or higher, and three variables meeting this criterion loaded on the second factor (i.e., slightly below the threshold of four or more variables per factor; see Beavers et al., 2013; Gaudagnli & Velicer, 1988). The first and second factors explained 47% and 12% of the shared variance, respectively.

The first factor comprised all eight writing quality ratings and one author characteristic rating (i.e., Intelligence). A second author characteristic (i.e., Knowledgeability) was just below the threshold for inclusion (loading = 0.58). This factor suggests two findings. First, raters were generally making a holistic evaluation of writing rather than distinct judgments of conventions, organization, fluency, and so on—perhaps indicating a halo effect (e.g., Engelhard, 1994; Knoch, Read, & Randow, 2007; and see Gansle, VanDerHeyden, Noell, Resetar, & Williams, 2006). Second, writers' intellectual abilities (i.e., intelligence and perhaps knowledgeability) seemed to also be embedded in judgments of writing quality.

The second factor comprised three author characteristics (i.e., Generosity, Kindness, and Loyalty), all of which were interpersonal rather than intellectual. No writing quality ratings loaded on this factor. This pattern suggests that judgments of authors' *personality* were distinct from judgments of writing or intellectual *ability* when evaluating texts without lower- or higher-level errors. However, the two factors were moderately correlated, $r = .46$, $p < .05$, suggesting that although writing and personality judgments were separable, they did influence one another. Writing perceived as lower quality may have led to harsher judgments of authors' generosity, kindness, and loyalty. Conversely, perhaps perceptions of the author as unkind or disloyal (e.g., stemming from reactions to essay content) led raters to be more critical of the writing. Although a correlation does not permit a clear determination, the former interpretation seems more likely.

Table 5
Model Fit Results for the Four Exploratory Factor Analyses

Error Type	Recommended Factor Solution			Model Fit Statistics		
	Kaisers Criterion	Scree Plot	Parallel Analysis	χ^2 (df)	RMSEA [CI95%]	SRMR
No Errors						
1 factor				275.75*** (104)	0.15 [0.13, 0.18]	0.11
2 factors			**2 factors**	**207.42*** (89)**	**0.14 [0.11, 0.16]**	**0.07**
3 factors	3 factors	3 factors		169.20*** (75)	0.13 [0.11, 0.16]	0.06
4 factors				169.61*** (62)	0.16 [0.13, 0.19]	0.04
5 factors				84.31** (50)	0.10 [0.06, 0.14]	0.03
Low-Level						
1 factor				269.92*** (104)	0.15 [0.13, 0.17]	0.07
2 factor	**2 factors**	**2 factors**	**2 factors**	**204.03*** (89)**	**0.14 [0.11, 0.16]**	**0.05**
3 factor				160.38*** (75)	0.13 [0.10, 0.16]	0.04
4 factor				90.36* (62)	0.08 [0.04, 0.12]	0.02
5 factor				99.32*** (50)	0.12 [0.08, 0.15]	0.02
High-Level						
1 factor				257.71*** (104)	0.15 [0.12, 0.17]	0.08
2 factor	**2 factors**	**2 factors**	**2 factors**	**148.72*** (89)**	**0.10 [0.07, 0.13]**	**0.04**
All Errors						
1 factor			1 factor	248.61*** (104)	0.14 [0.12, 0.16]	0.07
2 factors	**2 factors**	**2 factors**		**169.97*** (89)**	**0.11 [0.09, 0.14]**	**0.04**
3 factors				135.18** (75)	0.11 [0.08, 0.14]	0.04
4 factors				125.14*** (62)	0.12 [0.09, 0.15]	0.03

Note. *N* = 70. CI95% = 95% confidence interval for the RMSEA. Bold font indicates selected factor solution.
****p* < .001; ** *p* < .01; **p* < .05.

Writing and Author Ratings for 'Low-Level Errors Only' Essays

Sampling metrics indicated that it was appropriate to proceed with EFA. The KMO statistic was .93 ('excellent'), and initial communalities of all variables were greater than or equal to .60.

Models with one to six factors were considered. All but the six-factor model converged, which was omitted (see Table 5). After considering model fit criteria, a two-factor model was selected as optimal. Specifically, Kaiser's criterion and the scree plot indicated a two-factor model, although the parallel analysis favored a one-factor model. Inspection of the factor loadings also led to favoring a one or two-factor model. All five models yielded statistically significant chi-square values, indicating some misspecification to the model. The RMSEA dropped below the 0.10 threshold for the four-factor model. The SRMR became acceptable starting with the two-factor model and decreased with each successive model. The other models had insufficient numbers of appreciable loadings on one or more variables to warrant a meaningful factor.

The pattern matrix for the two-factor solution is presented in Table 6. The first factor had nine loadings of 0.60 or higher, and the second factor had three loadings that met this criterion (i.e., slightly below the threshold of four or more variables per factor). The first and second factors explained 64% and 8% of the shared variance among observed variables, respectively. The first factor comprised seven writing quality ratings (i.e., all traits except Conventions) and two author characteristics (i.e., Intelligence and Knowledgeability). One additional author characteristic was slightly below threshold (i.e., Creativity, loading = 0.56). The second factor comprised three interpersonal author characteristics (i.e., Generosity, Kindness, and Loyalty).

Overall, these findings largely replicate patterns from the No Errors essays. When evaluating texts with low-level errors only, raters seemed to make a holistic judgment of writing quality that incorporated the authors'

Table 6
Exploratory Factor Analysis Pattern Matrices for Essay Error Types

Rating	Essay Error Pattern							
	No Errors		Low-Level Errors		High-Level Errors		All Errors	
	Factor 1	Factor 2	Factor 1	Factor 2	Factor 1	Factor 2	Factor 1	Factor 2
Writing Traits								
Conventions	**0.60***	0.16	0.49*	0.16	0.49*	0.23	**0.73***	-0.05
Enjoyable	**0.83***	-0.06	**0.94***	-0.10	**0.63***	0.32*	**0.77***	0.01
Ideas and Content	**0.83***	-0.06	**0.90***	-0.08	**0.69***	0.24	**0.85***	0.07
Organization	**0.81***	0.00	**0.81***	0.03	**0.61***	0.23	**0.92***	-0.11
Persuasiveness	**0.69***	-0.12	**0.96***	-0.11	**0.84***	0.07	**0.95***	-0.02
Sentence Fluency	**0.85***	0.01	**0.81***	-0.07	**0.91***	-0.09	**0.74***	0.07
Voice	**0.85***	0.04	**0.73***	0.12	**0.88***	-0.15	**0.63***	0.20
Word Choice	**0.69***	0.12	**0.62***	0.26	**0.86***	-0.01	**0.81***	0.05
Author Traits								
Creativity	0.40*	0.24	0.56*	0.33	0.28	0.59*	0.28	0.56*
Generosity	0.01	**0.87***	0.04	**0.91***	0.09	**0.72***	0.03	**0.86***
Hard-working	0.38*	0.29	0.48*	0.40	0.15	**0.77***	0.41	0.48*
Intelligence	**0.65***	0.15	**0.63***	0.31	0.56*	0.38*	0.35	0.55*
Kindness	-0.03	**0.86***	-0.04	**0.88***	0.00	**0.77***	-0.15*	**0.94***
Knowledgeability	0.58*	0.21	**0.65***	0.23	0.17	**0.72***	0.57*	0.37
Loyalty	-0.04	**0.66***	0.14	**0.72***	-0.23*	**0.92***	0.04	**0.85***
Thoughtfulness	0.36*	0.42*	0.52*	0.35	0.17	**0.73***	0.29	**0.61***
Variance Explained	47%	12%	64%	8%	62%	10%	66%	8%

Note. *N* = 70. Factors rotated using Oblimin rotation. Appreciable factor loadings (⩾ 0.60) are indicated using bold font. $^{*}p < .05$.

intellectual abilities, and seemed to make a separate judgment of authors' personality. The two factors were again moderately correlated, $r = .64$, $p < .05$, and the relationship was stronger than with No Errors essays.

Writing and Author Ratings for 'High-Level Errors Only' Essays

Sampling metrics indicated that it was appropriate to proceed with EFA. The KMO statistic was .89 ('very good'), and initial communalities for 15 of the 16 variables were greater than or equal to .60.

Only the one- and two-factor models converged (see Table 5); models of three or more factors were thus omitted. After considering model fit criteria, a two-factor model was selected as optimal. Kaiser's criterion, the scree plot, and the parallel analysis all favored a two-factor model. Both models yielded a significant chi-square statistic, indicating some misspecification to the model. The RMSEA for the one-factor model was 0.15, indicating a poor-fitting model. The RMSEA of the two-factor model yielded a minimally acceptable value of 0.10. The SRMR for the two-factor solution was 0.04, indicating adequate model fit.

The pattern matrix for the two-factor solution is given in Table 6. The first factor had seven loadings of 0.60 or higher, and the second factor had six loadings that met this criterion. The first and second factors explained 62% and 10% of the shared variance among observed variables, respectively. The first factor comprised seven writing traits (i.e., all except Conventions) and no author characteristics, although Intelligence was just below threshold (loading = 0.56). The second factor comprised six author traits (i.e., Generosity, Hard-working, Kindness, Knowledgeability, Loyalty, and Thoughtfulness), and Creativity was just below threshold (loading = 0.59).

These patterns both corroborate and diverge from prior findings. First, raters again seemed to make broad writing quality evaluations and author personality judgments that were distinct but related ($r = .64$, $p < .05$), with a similar magnitude as Low-Level Errors Only essays. Although not significant, Intelligence somewhat loaded on the writing quality factor but not the author personality factor—again suggesting that perceived intellectual abilities play a role in writing judgments. However, the author personality factor included many more characteristics in this model. In addition to interpersonal traits like being generous or kind, this factor also included work ethic and thoughtfulness along with possible intellectual traits.

The critical difference between this model and prior models was the presence of higher-level writing errors, such as disorganization, missing arguments, and illogical examples. Compared to essays exhibiting no errors or only lower-level errors (e.g., spelling and grammar), raters seemed to make more sweeping evaluations of the authors themselves.

Writing and Author Ratings for 'All Errors' Essays

Sampling metrics indicated that it was appropriate to proceed with EFA. The KMO statistic was .92 ('excellent'), and initial communalities of all variables were greater than or equal to .60.

Models for the one- through four-factor solutions converged, but not the five-factor and six-factor solutions, which were omitted (see Table 5). Based on review of model fit metrics, the two-factor solution was retained as the optimal model. Kaiser's criterion and the scree plot recommended the two-factor solution although the parallel analysis pointed to a one-factor solution. Inspection of factor loadings revealed that the three-factor solution only had one appreciable loading (i.e., a loading ⩾ .60) on the third factor, and the four-factor solution had no appreciable loadings on the fourth factor. All models yielded a significant chi-square statistic, indicating some misspecification to the model. RMSEA values presented similar information. In all cases, the RMSEA exceeded 0.10, although the lower bound of the 95% confidence interval for the two-factor solution dropped below this threshold. The SRMR for the two-factor solution was 0.04, indicating adequate fit.

The pattern matrix for the two-factor solution is presented in Table 6. The first factor had eight loadings of 0.60 or higher, and the second factor had four loadings that met this criterion. The first and second factors explained 66% and 8%, respectively, of the shared variance among the observed variables. The first factor comprised all eight writing quality traits and no author characteristic ratings, although Knowledgeability was just below threshold (loading = 0.57). The second factor comprised four author characteristic ratings (i.e., Generosity, Kindness, Loyalty, and Thoughtfulness), with two others near the threshold (i.e., Creativity, loading = 0.56; and Intelligence, loading = 0.55).

The pattern for All Errors essays was most similar to the model for High-Level Errors Only essays, although the model again corroborated prior results. Raters seemed to make a holistic judgment of writing quality that included elements of intellectual ability, and distinct but strongly related judgments of authors' personality ($r = .75$, $p < .05$). Essays that exhibited both kinds of errors also demonstrated the strongest correlation between factors. As above, the presence of higher-level writing errors seemed to result in somewhat broader evaluations of the authors—not just interpersonal traits (e.g., loyalty and kindness) but also effort and intellectual traits.

Discussion

For better or worse, writers are judged by their audience, and these evaluations encompass both the text and the authors themselves (Cox et al., 2017; Figueredo & Varnhagen, 2005; Authors, 2017; Vignovic & Thompson, 2010). Moreover, such assessments are consequential. Outside of school, employers may make hiring decisions based on writing skills or personal characteristics 'revealed' through one's writing (Hoover, 2013); teachers might make judgments about students' capabilities or conduct (e.g., Johnson & VanBrackle, 2012); and individuals may even make decisions about potential roommates (Boland & Queen, 2016). Although readers' perceptions of writing errors can be inaccurate, biased, and unfair, the impact of those perceptions cannot be disregarded.

The current paper follows on prior research (Johnson et al., 2017) to further investigate relationships between college students' judgments of writing and author as a function of perceived writing errors. Unanswered questions pertained to whether evaluations of writing (RQ1) and author (RQ2) represented holistic or nuanced judgments, the degree of overlap between writing and author assessment constructs (RQ3), and the influence of error patterns on observed latent constructs (RQ4).

Current results suggest that students tended to form holistic impressions of writing quality and authors rather than distinct judgments about individual traits. Analyses consistently generated models wherein most variables loaded on relatively few factors that explained over half the variance (59-74%). The first factor included all or most writing traits, and the second factor comprised multiple author characteristics. Student raters appeared to make holistic judgments rather than nuanced judgments. In practice, college students did not make distinct evaluations of conventions, persuasiveness, word choice, and so on. All of these variables were likely considered together to form a collective evaluation or a subset of traits dominated the perception of all others.

Results also suggest that there is subtle overlap in judgments of writing and author. In particular, the dominant construct in all models—tentatively labeled *Writing Quality and Skill*—tended to incorporate all or most writing traits along with several 'intellectual' author traits. Assessments of whether an essay was well-written were conflated with perceiving authors as smart or informed. A second construct, *Author Personality*, seemed to focus on writers' perceived 'interpersonal' traits (e.g., generosity, kindness, and loyalty) with little to no contribution from writing quality. In contrast to intellectual abilities, interpersonal qualities seemed to be judged separately from writing quality. In all cases, however, these two constructs were significantly and positive correlated (*r*s > .60). Thus, although the two judgments are separable, they very likely influence each other. In accord with prior research, higher writing quality may lead students to perceive their peers as more giving, friendly, or trustworthy; and favorable beliefs or biases toward the author based on the content of their essay may lead to more forgiving review of the essay.

A subtle influence of error pattern was perhaps observed on latent judgments of writing and author. The presence of higher-level writing errors seemed to trigger more sweeping (or less focused) judgments of author personality and to attenuate the connection between writing quality and intellectual traits. This effect was most noticeable for essays that exhibited *only* higher-level errors of disorganization, missing arguments, and illogical evidence. In this case, the Author Personality construct comprised Generosity, Hard-working, Kindness, Knowledgeability, Loyalty, and Thoughtfulness, whereas for other error patterns the contributing variables focused on interpersonal traits. One possibility suggested by the data is that higher-level errors were not penalized to the same degree as lower-level errors. When rating such essays, higher-level errors may have been perceptible but less salient, which also reduced the salience of personality judgments. Consequently, student raters provided more global or vague evaluations of the author.

There were several limitations to the current study that should be addressed in future research. First, the sample size was relatively small for conventional EFA analyses. Although the within-subjects design was strong for assessing the effects of different error patterns on student raters' perceptions (i.e., the purpose of the original study, Johnson et al., 2017), a sample of 70 participants was somewhat low for conducting EFAs. Multiple checks of data adequacy for EFA were implemented, including the KMO statistic, number of variables with communalities greater than .60, and number of factor loadings per factor. Importantly, for all analyses, these checks indicated that the sample was acceptable. Nonetheless, it would be worthwhile for future modeling to build on current findings using larger samples. Future studies also represent an opportunity to recruit more diverse samples or introduce other manipulations (see below) to further explore perceptions of errors, writing quality, and authors.

Other limitations pertained to the essay stimuli. First, for consistency, all essays were argument essays constructed in response to the same prompt about 'patience.' This design did not afford testing for prompt-based effects, such as whether certain topics draw students' attention to technical aspects of writing quality or to characteristics of the authors. Similarly, this design does not permit exploration of genre effects. For example, an argument essay about the value of patience is likely to seem more personal in nature than an expository text about a scientific phenomenon. Thus, judgments about authors' personality may have been more salient in this study. In future work, it will be useful to manipulate the genre of the essay stimuli (e.g., argument, expository, or narrative) along with the presence or absence of self-references (e.g., first-person pronouns and anecdotes).

It is worth noting, however, that prior research on perceptions of writing and authors have been conducted with a variety of writing types ranging from formal classroom assignments to informal emails.

Implications for Peer Assessment

Given that evaluations of writing, errors, and authors can be conflated, one set of implications for peer assessment of writing pertains to how these effects might be mitigated or under what conditions they are exacerbated. Students are often skeptical of peer assessment and express doubts that their peers are capable of performing reliable and valid assessments, particularly when course grades are at stake (Gielen, Peeters, Dochy, Onghena, & Struyven, 2010; Kaufman & Schunn, 2011; van Zundert, Sluijsmans, & van Merriënboer, 2010). Students may be worried that their peers are forming personal or intellectual judgments and then biasing their reviews and scores based on these judgments. The results of the current study suggest that this is a plausible concern. If students recognize that they are making such judgments about other students, a reasonable interpersonal inference is that their peers are doing the same (see Panadero, 2016; Panadero et al., 2013; van Gennip, Segers, & Tillema, 2009).

Masked review policies are often instituted to avoid possible bias or unfairness in peer assessment (e.g., Kaufman & Schunn, 2011; Panadero & Alqassab, 2019) and scholarly publishing (e.g., Lee, Sugimoto, Zhang, & Cronin, 2013). In principle, by obscuring the identities or backgrounds of their peers (e.g., name, race, gender, or nationality), student assessors cannot use this information to offer biased assessments or interpretations. However, the essay rating task employed in this study was effectively blind—student raters were given no information about the supposed authors—yet text and author judgments were still overlapping or strongly correlated. Thus, 'blind review' did not solve the problem of making personal judgments about authors based on perceived writing errors.

A more direct approach may be to provide additional training to students that counteracts unwarranted inferences about their peers (e.g., Goodwin, 2016; May, 2008; Soltero-González, Escamilla, & Hopewell, 2012) and improves writing assessment literacy (see Crusan, Plakans, & Gebril, 2016; Weigle, 2007). Traditionally, expert raters are trained and assessed based on inter-rater agreement (e.g., Huot, 1990; Jonsson & Svingby, 2007), yet such agreement does not guarantee a lack of bias. Instead of disregarding inferences about author characteristics and errors, high agreement could simply indicate that raters are making similar interpretations (e.g., conflating writing skill and intellectual ability). Thus, training that explicitly tackles correspondence bias or other social-perceptual biases (e.g., May, 2008) may be particularly beneficial for student raters who lack writing knowledge or proficiency. For instance, training focused on perspective-taking (i.e., considering the perspectives of others in terms of point of view, location, or time) has been shown to reduce the occurrence of the fundamental attribution error among adults (e.g., Hooper, Erdogan, Keen, Lawton, & McHugh, 2015). One avenue for future research may thus be to incorporate perspective-taking exercises into peer assessment training. Students can be taught to be more mindful of how written products may not reflect the true circumstances or identity of the writer (e.g., frequent typos may reflect writing under high time pressure rather than 'laziness' or 'lack of intelligence').

In addition, research on rubrics has shown that they can improve peer (and self) assessment validity and reliability (Jonnson & Svingby, 2007; Panadero, Romero, & Strijbos, 2013; Panadero & Jonsson, 2013). In a recent meta-analysis, students who were trained to provide ratings demonstrated greater learning gains than those who completed the peer assessment without training (Li et al., 2019). Notably, the current study did not employ detailed assessment rubrics or rubric-referenced training. Participants were provided with brief descriptions of eight writing traits and eight author traits (see Table 1), but were not given detailed criteria or benchmark examples. This method was implemented because the aim was to assess perceptions of errors rather than adherence to a rubric or checklist. However, although the traits and terms were fairly straightforward, participants likely possessed differential understanding of the concepts (e.g., epistemological beliefs about 'intelligence' and 'knowledge' or personal experiences with 'generosity' and 'kindness'). In future research, a plausible hypothesis is that rubric-based training would result in more distinct trait judgments—the underlying factor structure might exhibit a larger number of latent assessment constructs rather than a few holistic constructs. It is unclear whether such training would reduce, exacerbate, or have no effect on the occurrence of personal author judgments or the connections between writing and author evaluations. To further explore these outcomes, rubric-based approaches might be further enhanced via concrete strategies for avoiding personal judgments about authors when assessing writing or writing errors. Rubrics and exemplars could not only clarify the meaning of ‘conventions,’ ‘sentence fluency,’ ‘loyalty,’ or ‘creativity,’ but also establish criteria for when judgments about such traits are (or are not) warranted.

Conclusion

Writing skills are critical for success in academic, professional, and social settings. Although a great deal of attention is paid to teaching writing and evaluating writing products in reliable and valid ways, current research suggests that focus should also be directed to underlying relationships among perceptions of writing and writers. Moderate to strong links were observed between ratings of ‘writing quality and skill’ and ‘author personality,’ and these relationships were strengthened in the presence of perceived writing errors. It makes sense that writing errors could or should have a valid impact on writing quality judgments. However, the spillover onto perceptions of authors’ personal characteristics may be representative of latent biases, perhaps stemming from differences in education, identity, culture, and so on. As the stakes for writing performance increase, it is important for assessors and policymakers to take steps to recognize and mitigate these effects.

Acknowledgments

This research was supported in part by a Fulton Undergraduate Research Initiative award to Adam C. Johnson from the Ira A. Fulton Schools of Engineering at Arizona State University.

Conflict of Interest

The authors declare that they have no conflict of interests.

References

Ackerman, J. M. (1993). The promise of writing to learn. *Written Communication, 10*(3), 334–370. https://doi.org/10.1177/0741088393010003002

Attali, Y. (2016). A comparison of newly-trained and experienced raters on a standardized writing assessment. *Language Testing, 33*, 99–115. https://doi.org/10.1177/0265532215582283

Arnold, K. M., Umanath, S., thio, K., Reilly, W. B., McDaniel, M. A., & Marsh, E. J. (2017). Understanding the cognitive processes involved in writing to learn. *Journal of Experimental Psychology: Applied*, *23*(2), 115–127. https://doi.org/10.1037/xap0000119

Bangert-Drowns, R. L., Hurley, M. M., & Wilkinson, B. (2004). The effects of school-based writing-to-learn interventions on academic achievement: A meta-analysis. *Review of Educational Research*, *74*(1), 29–58. https://doi.org/10.3102/00346543074001029

Bauman, C. W., & Skitka, L. J. (2010). Making attributions for behaviors: The prevalence of correspondence bias in the general population. *Basic and Applied Social Psychology*, *32*, 269–277. https://doi.org/10.1080/01973533.2010.495654

Beavers, A. S., Lounsbury, J. W., Richards, J. K., Huck, S. W., Skolits, G. J., & Esquivel, S. L. (2013). Practical considerations for using exploratory factor analysis in educational research. *Practical Assessment, Research & Evaluation, 18*(6), 1–13. https://scholarworks.umass.edu/pare/vol18/iss1/6

Boland, J. E., & Queen, R. (2016). If you’re house is still available, send me an email: personality influences reactions to written errors in email messages. *PloS One*, *11*(3), e0149885. https://doi.org/10.1371/journal.pone.0149885

Breland, H. M., & Jones, R. J. (1982). Perceptions of writing skill. *Written Communication, 1*(1), 101–119. https://doi.org/10.1177/0741088384001001005

Brown, T. A. (2006). *Confirmatory factor analysis for applied research.* Guilford.

Byrne, B. M. (2013). *Structural equation modeling with Mplus: Basic concepts, applications, and programming.* Routledge.

Cho, K., & MacArthur, C. (2011). Learning by reviewing. *Journal of Educational Psychology*, *103*, 73–84. https://doi.org/10.1037/a0021950

Cho, K., Schunn, C. D., & Charney, D. (2006). Commenting on writing: Typology and perceived helpfulness of comments from novice peer reviewers and subject matter experts. *Written Communication*, *23*(3), 260–294. https://doi.org/10.1177/0741088306289261

Cox, D., Cox, J. G., & Cox, A. D. (2017). To err is human? How typographical and orthographical errors affect perceptions of online reviewers. *Computers in Human Behavior*, *75*, 245–253. https://doi.org/10.1016/j.chb.2017.05.008

Crossley, S. A., Roscoe, R. D., & McNamara, D. S. (2014). What is successful writing? An investigation in the multiple ways writers can write successful essays. *Written Communication*, *31*(2), 184–214. https://doi.org/10.1177/0741088314526354

Crusan, D., Plakans, L., & Gebril, A. (2016). Writing assessment literacy: Surveying second language teachers' knowledge, beliefs, and practices. *Assessing Writing*, *28*, 43–56. https://doi.org/10.1016/j.asw.2016.03.001

Devitt, A. J. (2004). *Writing genres*. Southern Illinois University Press.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*, 93–112. https://doi.org/10.1111/j.1745-3984.1994.tb00436.x

Elton, L., (2010). Academic writing and tacit knowledge. *Teaching Higher Education*, *15*, 151–160. https://doi.org/10.1080/13562511003619979

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research*, *70*(3), 287–322. https://doi.org/10.3102/00346543070003287

Fathi, J. & Khodabakhsh, M. R. (2019). The role of self-assessment and peer-assessment in improving writing performance of Iranian EFL students. *International Journal of English Language & Translation Studies, 7*(3), 1–10.

Field, A. (2013). *Discovering statistics using IBM SPSS Statistics* (4th ed.). Sage.

Figueredo, L., & Varnhagen, C. K. (2005). Didn't you run the spell checker? Effects of type of spelling error and use of a spell checker on perceptions of the author. *Reading Psychology*, *26*(4-5), 441–458. https://doi.org/10.1080/02702710500400495

Gansle, K. A., VanDerHeyden, A. M., Noell, G. H., Resetar, J. L., & Williams, K. L. (2006). The technical adequacy of curriculum-based and rating-based measures of written expression for elementary school students. *School Psychology Review*, *35*, 435–450.

Gao, Y., Schunn, C. D., & Yu, Q. (2019) The alignment of written peer feedback with draft problems and its impact on revision in peer assessment, *Assessment & Evaluation in Higher Education, 44*(2), 294–308. https://doi.org/10.1080/02602938.2018.1499075

Gielen, S., Peeters, E., Dochy, F., Onghena, P., & Struyven, K. (2010). Improving the effectiveness of peer feedback for learning. *Learning and Instruction*, *20*, 304–315. https://doi.org/10.1016/j.learninstruc.2009.08.007

Gielen, S., Tops, L., Dochy, F., Onghena, P., & Smeets, S. (2010). A comparative study of peer and teacher feedback and of various peer feedback forms in a secondary school writing curriculum. *British Educational Research Journal*, *36*, 143–162. https://doi.org/10.1080/01411920902894070

Godley, A., & Escher, A. (2012). Bidialectual African American adolescents' beliefs about spoken language expectations in English classrooms. *Journal of Adolescent and Adult Literacy*, *55*(8), 704–713. https://doi.org/10.1002/JAAL.00085

Goodwin, S. (2016). A Many-Facet Rasch analysis comparing essay rater behavior on an academic English reading/writing test used for two purposes. *Assessing Writing*, *30*, 21–31. https://doi.org/10.1016/j.asw.2016.07.004

Guadagnoli, E. & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin, 103*, 265–275. https://doi.org/10.1037/0033-2909.103.2.265

Hacker, D., & Sommers, N. (2016). *Rules for writers* (8th ed.). Bedford/St. Martin's.

Hoover, B. (2013, March 4th). Good grammar should be everyone's business. *Harvard Business Review*. https://hbr.org/2013/03/good-grammar-should-be-everyon

Huot, B. (1990). Reliability, validity, and holistic scoring: What we know and what we need to know. *College Composition and Communication*, *41*(2), 201–213. https://doi.org/10.2307/358160

Hyland, K. (2007). Genre pedagogy: Language, literacy, and L2 writing instruction. *Journal of Second Language Writing*, *16*(3), 148–164. https://doi.org/10.1016/j.jslw.2007.07.005

Jeong, A., Li, H., & Pan, A. J. (2017). A sequential analysis of responses in online debates to postings of students exhibiting high versus low grammar and spelling errors. *Educational Technology Research and Development*, *65*(5), 1175–1194. https://doi.org/10.1007/s11423-016-9501-2

Johnson, A. C., Wilson, J., & Roscoe, R. D. (2017). College student perceptions of writing errors, text quality, and author characteristics. *Assessing Writing*, *34*, 72–87.

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 example. *Assessing Writing*, *17*, 35–54. https://doi.org/10.1016/j.asw.2011.10.001

Jonnson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, *2*(2), 130–144. https://doi.org/10.1016/j.edurev.2007.05.002

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika, 39*, 31–36.

Kaufman, J. H., & Schunn, C. D. (2011). Students' perceptions about peer assessment for writing: Their origin and impact on revision work. *Instructional Science*, *39*(3), 387–406. https://doi.org/10.1007/s11251-010-9133-6

Kenny, D. A. (2014, February 6). *Measuring model fit*. http://davidakenny.net/cm/fit.htm

Klein, P. D. (1999). Reopening inquiry into cognitive processes in writing-to-learn. *Educational Psychology Review, 11*(3), 203–270. https://doi.org/10.1023/A:1021913217147

Knoch, U., Read, J., & Randow, J. (2007). Re-training writing raters online: How does it compare to face-to-face training? *Assessing Writing*, *12*, 26–43. https://doi.org/10.1016/j.asw.2007.04.001

Kreiner, D. S., Schnakenberg, S. D., Green, A. G., Costello, M. J., & McClin, A. F. (2002). Effects of spelling errors on the perception of writers. *The Journal of General Psychology*, *129*(1), 5–17. https://doi.org/10.1080/00221300209602029

Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in peer review. *Journal of the Association for Information Science and Technology*, *64*(1), 2–17. https://doi.org/10.1002/asi.22784

Li, H., Xiong, Y., Hunter, C. V., Guo, X., & Tywoniw, R. (2019) Does peer assessment promote student learning? A meta-analysis, *Assessment & Evaluation in Higher Education*. https://doi.org/10.1080/02602938.2019.1620679

Li, H., Xiong, Y., Zang, X., Kornhaber, M. L., Lyu, Y., Chung, K. S., & K. Suen, H. (2016). Peer assessment in the digital age: A meta-analysis comparing peer and teacher ratings. *Assessment & Evaluation in Higher Education*, *41*(2), 245–264. https://doi.org/10.1080/02602938.2014.999746

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing*, *28*, 543–560. https://doi.org/10.1177/0265532211406422

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*, 84–99. https://doi.org/10.1037/1082-989X.4.1.84

Magnifico, A. M. (2010). Writing for whom? Cognition, motivation, and a writer's audience. *Educational Psychologist*, *45*(3), 167–184. https://doi.org/10.1080/00461520.2010.493470

May, G. L. (2008). The effect of rater training on reducing social style bias in peer evaluation. *Business Communication Quarterly*, *71*(3), 297–313. https://doi.org/10.1177/1080569908321431

McNamara, D. S. (2013). The epistemic stance between the author and reader: A driving force in the cohesion of text and writing. *Discourse Studies*, *15*(5), 579–595. https://doi.org/10.1177/1461445613501446

Morin-Lessard, E., & McKelvie, S. J. (2019). Does writeing rite matter? Effects of textual errors on personality trait attributions. *Current Psychology*, *38*(1), 21–32. https://doi.org/10.1007/s12144-017-9582-z

Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus user's guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.

Olinghouse, N. G., Graham, S., & Gillespie, A. (2014). The relationship of discourse and topic knowledge to fifth graders' writing performance. *Journal of Educational Psychology*, *101*, 37–50. https://doi.org/10.1037/a0037549

Panadero, E. (2016). Is it safe? Social, interpersonal, and human effects of peer assessment. In G. T. L., Brown, & L. R. Harris (Eds.), *Handbook of social and human conditions in assessment* (pp. 247–266). Routledge.

Panadero, E., & Alqassab, M. (2019). An empirical review of anonymity effects in peer assessment, peer feedback, peer review, peer evaluation and peer grading. *Assessment and Evaluation in Higher Education*, *44*(8), 1253–1278. https://doi.org/10.1080/02602938.2019.1600186

Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review*, *9*, 129–144. https://doi.org/10.1016/j.edurev.2013.01.002

Panadero, E., Romero, M., & Strijbos, J. (2013). The impact of a rubric and friendship on peer assessment: Effects on construct validity, performance, and perceptions of fairness and comfort. *Studies in Educational Evaluation*, *39*(4), 195–203. https://doi.org/10.1016/j.stueduc.2013.10.005

Patchan, M. M., Charney, D., & Schunn, C. D. (2009). A validation study of students' end comments: Comparing comments by students, a writing instructor, and a content instructor. *Journal of Writing Research*, *1*(2), 124–152.

Patchan, M. M., Hawk, B., Stevens, C. A., & Schunn, C. D. (2013). The effects of skill diversity on commenting and revisions. *Instructional Science*, *41*(2), 381–405. https://doi.org/10.1007/s11251-012-9236-3

Patchan, M. M., & Schunn, C. D. (2015). Understanding the benefits of providing peer feedback: How students respond to peers' texts of varying quality. *Instructional Science*, *43*(5), 591–614. https://doi.org/10.1007/

s11251-015-9353-x

Patchan, M. M., & Schunn, C. D. (2016). Understanding the effects of receiving peer feedback for text revision: Relations between author and reviewer ability. *Journal of Writing Research*, *8*(2), 227–265. https://doi.org/10.17239/jowr-2016.08.02.03

Patchan, M. M., Schunn, C. D., & Clark, R. J. (2011). Writing in the natural sciences: Understanding the effects of different types of reviewers on the writing process. *Journal of Writing Research*, *2*(3), 365–393.

Patchan, M. M., Schunn, C. D., & Clark, R. J. (2018). Accountability in peer assessment: Examining the effects of reviewing grades on peer ratings and peer feedback. *Studies in Higher Education*, *43*(12), 2263–2278. https://doi.org/10.1080/03075079.2017.1320374

Ramon-Casas, M., Nuño, N., Pons, F., & Cunillera, T. (2019). The different impact of a structured peer-assessment task in relation to university undergraduates' initial writing skills. *Assessment & Evaluation in Higher Education, 44*(5), 653–663. https://doi.org/10.1080/02602938.2018.1525337

Reiff, M. J., & Bawarshi, A. (2011). Tracing discursive resources: How students use prior genre knowledge to negotiate new writing contexts in first-year composition. *Written Communication*, *28*(3), 312–337. https://doi.org/10.1177/0741088311410183

Roscoe, R. D. (2014). Self-monitoring and knowledge-building in learning by teaching. *Instructional Science*, *42*, 327–251. https://doi.org/10.1007/s11251-013-9283-4

Ruscio, J., & Roche, B. (2012). Determining the number of factors to retain in an exploratory factor analysis using comparison data of known factorial structure. *Psychological Assessment, 24*, 282–292. https://doi.org/10.1037/a0025697

Schunn, C., Godley, A., & DeMartino, S. (2016). The reliability and validity of peer review of writing in high school AP English classes. *Journal of Adolescent and Adult Literacy*, *60*(1), 13–23. https://doi.org/10.1002/jaal.525

Sluijsmans, D. M. A., Brand-Gruwel, S., van Merriënboer, J. J. G., & Bastiaens, T. J. (2003). The training of peer assessment skills to promote the development of reflection skills in teacher education. *Studies in Educational Evaluation*, *29*(1), 23–42. https://doi.org/10.1016/S0191-491X(03)90003-4

Soltero-González, L., Escamilla, K., & Hopewell, S. (2012). Changing teachers' perceptions about the writing abilities of emerging bilingual students: Towards a holistic bilingual perspective on writing assessment. *International Journal of Bilingual Education and Bilingualism*, *15*(1), 71–94. https://doi.org/10.1080/13670050.2011.604712

Steiger, J. H., & Lind, J. C. (1980, May). *Statistically based tests for the number of common factors*. Paper presented at the annual Spring Meeting of the Psychometric Society, Iowa City, IA.

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research*, *68*, 249–276. https://doi.org/10.2307/1170598

Topping, K. (2009). Peer assessment. *Theory into Practice*, *48*, 20–27. https://doi.org/10.1080/00405840802577569

Topping, K. J., Smith, E. F., Swanson, I., & Elliot, A. (2000). Formative peer assessment of academic writing between postgraduate students. *Assessment & Evaluation in Higher Education*, *25*(2), 149–169. https://doi.org/10.1080/713611428

van Gennip, N. A. E., Segers, M. S. R., & Tillema, H. H. (2009). Peer assessment for learning from a social perspective: The inference of interpersonal variables and structural features. *Educational Research Review*, *4*(1), 41–54. https://doi.org/10.1016/j.edurev.2008.11.002

van Zundert, M., Sluijsmans, D., & van Merriënboer, J. (2010). Effective peer assessment processes: Research findings and future directions. *Learning and Instruction*, *20*, 270–279. https://doi.org/10.1016/j.learninstruc.2009.08.004

Varner (Allen), L. K., Roscoe, R. D., & McNamara, D. S. (2013). Evaluative misalignment of 10th-grade student and teacher criteria for essay quality: An automated textual analysis. *Journal of Writing Research*, *5*(1), 35–59. https://doi.org/10.17239/jowr-2013.05.01.2

Vignovic, J. A., & Thompson, L. F. (2010). Computer-mediated cross-cultural collaboration: Attributing communication errors to the person versus the situation. *Journal of Applied Psychology*, *95*(2), 265–276. https://doi.org/10.1037/a0018628

Wang, J., & Engelhard, G. (2019). Conceptualizing rater judgments and rating processes for rater-mediated assessments. *Journal of Educational Measurement, 56*, 582–602. https://doi.org/10.1111/jedm.12226

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*, 263– 287. https://doi.org/10.1177/026553229801500205

Weigle, S. C. (2007). Teaching writing teachers about assessment. *Journal of Second Language Writing*, *16*(3), 194–209. https://doi.org/10.1016/j.jslw.2007.07.004