

Rapid #: -19446106

CROSS REF ID: **1851893**

LENDER: **NNY :: Main Library**

BORROWER: **AZS :: Main Library**

TYPE: Article CC:CCG

JOURNAL TITLE: Measurement

USER JOURNAL TITLE: Measurement : interdisciplinary research and perspectives.

ARTICLE TITLE: Exploring Rater Accuracy Using Unfolding Models Combined with Topic Models: Incorporating Supervised Latent Dirichlet Allocation

ARTICLE AUTHOR: Wheeler, Jordan M

VOLUME: 20

ISSUE: 1

MONTH:

YEAR: 2022

PAGES: 34-46

ISSN: 1536-6367

OCLC #: 47774715

Processed by RapidX: 8/26/2022 12:26:01 PM

This material may be protected by copyright law (Title 17 U.S. Code)



Measurement: Interdisciplinary Research and Perspectives

ISSN: (Print) (Online) Journal homepage: <https://www.tandfonline.com/loi/hmes20>

Exploring Rater Accuracy Using Unfolding Models Combined with Topic Models: Incorporating Supervised Latent Dirichlet Allocation

Jordan M. Wheeler, George Engelhard & Jue Wang

To cite this article: Jordan M. Wheeler, George Engelhard & Jue Wang (2022) Exploring Rater Accuracy Using Unfolding Models Combined with Topic Models: Incorporating Supervised Latent Dirichlet Allocation, Measurement: Interdisciplinary Research and Perspectives, 20:1, 34-46, DOI: [10.1080/15366367.2021.1915094](https://doi.org/10.1080/15366367.2021.1915094)

To link to this article: <https://doi.org/10.1080/15366367.2021.1915094>



Published online: 02 Feb 2022.



Submit your article to this journal [↗](#)



Article views: 80



View related articles [↗](#)



View Crossmark data [↗](#)



Exploring Rater Accuracy Using Unfolding Models Combined with Topic Models: Incorporating Supervised Latent Dirichlet Allocation

Jordan M. Wheeler ^a, George Engelhard ^a, and Jue Wang ^b

^aEducational Psychology, The University of Georgia; ^bEducational and Psychological Studies, The University of Miami

ABSTRACT

Objectively scoring constructed-response items on educational assessments has long been a challenge due to the use of human raters. Even well-trained raters using a rubric can inaccurately assess essays. Unfolding models measure rater's scoring accuracy by capturing the discrepancy between criterion and operational ratings by placing essays on an unfolding continuum with an ideal-point location. Essay unfolding locations indicate how difficult it is for raters to score an essay accurately. This study aims to explore a substantive interpretation of the unfolding scale based on a supervised Latent Dirichlet Allocation (sLDA) model. We investigate the relationship between latent topics extracted using sLDA and unfolding locations with a sample of essays ($n = 100$) obtained from an integrated writing assessment. Results show that (a) three latent topics moderately explain ($r^2 = 0.561$) essay locations defined by the unfolding scale and (b) failing to use and/or cite the source articles led to essays that are difficult-to-score accurately.

KEYWORDS

Rater-mediated assessments; supervised latent Dirichlet allocation; topic models; unfolding model; hyperbolic cosine accuracy model

Rater-mediated assessments contain responses that require rater scoring, such as extended-response items, performance tasks, and portfolios. These assessments often allow students to better demonstrate their competencies, and are frequently used to assess high order thinking skills. Even if raters are thoroughly trained and follow a rubric, these rater-mediated assessments are still susceptible to potential biases, ranging from social positioning biases, such as gender, race, and class, to systematic biases, such as severity or leniency, and central tendency (Engelhard, 1994; Read et al., 2005; Saal et al., 1980). Nonetheless, these potential biases add a source of variance to the ratings that are irrelevant to the construct being measured and infringe on the validity and fairness of the ratings (Eckes, 2005; Rezaei & Lovorn, 2010).

Detecting rater biases can be achieved through various statistical methods (Aubin et al., 2018; Myford & Wolfe, 2003; Wind & Engelhard, 2012). These methods rely on a model-data fit approach that identify raters that exhibit aberrant response patterns. Although quantitative approaches are able to identify biases in rater-mediated assessments, further analyses are necessary for an evaluation of the causes for raters to elicit biases toward particular essays. A method that is able to identify and evaluate sources of inaccuracy would help policymakers prescribe various interventions to monitor and mitigate rater biases.

Conceptualizing rater judgments and the rating process, however, is a challenge. Previous research developed different models for examining rater effects, including latent trait modeling (Engelhard, 1994; Wolfe & McVay, 2012), hierarchical rater model (Patz et al., 2002), rater Bundle Model (Wilson & Hoskens, 2001), and generalized rater model (W. C. Wang et al., 2014). Recent studies have tried to understand this process through an unfolding model (J. Wang & Engelhard, 2019a, 2019b; J. Wang et al., 2016). These studies examined individual differences among raters in scoring the essays. For

instance, raters may show different levels of accuracy toward scoring the same essay. In other words, an essay may appear to be difficult-to-score across different raters. Therefore, instead of assuming that raters score the essays in a consistent manner, unfolding models examine individual differences among raters in scoring essays and help us understand to what degree raters assess essays differently. When the focus is placed onto the evaluation of rater accuracy, unfolding models define the essays based on their difficulty-to-score for individual raters.

A challenge when using an unfolding model to understand rater judgments is that the substantive interpretation of unfolding scale is not easily defined. J. Wang and Engelhard (2019b) used essay feature indices from Coh-Metrix to explore the meaning of an unfolding scale and found promising results. A quantitative analysis of the content of the essays through the use of topic models, which are a set of statistical models used to analyze textual data, could provide additional interpretations about the meaning of locations along the unfolding scale. The goal of topic models is to estimate latent topics found in a collection of essays. Latent topics are similar to principle components and represent clusters of words that have similar context across a collection of essays. Latent topics are used to infer the relationship between essays and words. A recent study showed that the results from topic models provide additional information beyond the score and provide insight into the writing process of students (Cardozo-Gaibisso et al., 2020). Topic models have also been used to analyze the effects from instructional writing interventions by identifying latent topics associated to different language use in students' essays (Duong et al., 2019; Kim et al., 2017). Therefore, topic models enable a robust analysis of textual data that may influence the scores assigned by raters to essays.

Purpose of the study

In this study, we used an unfolding model combined with a supervised topic model to evaluate essays written by students to an extended-response item in an integrated writing assessment. Integrated writing assessments require students to read source passages, and to utilize information from the passages in answering an essay prompt. Integrated writing assessments are becoming more common in statewide assessments. The unfolding model is used to define difficult-to-score essays, and the supervised topic model is used to estimate the latent topics from the content of the essays. The study investigates two questions: (1) what are the latent topics used across all essays? and (2) how do the latent topics relate to the unfolding locations of each essay? This information can be used to identify sources of inaccuracy based on the latent topics and ultimately improve rater training and the quality of ratings.

Methodology

Accuracy ratings

The accuracy ratings directly reflect the distance between the criterion and observed ratings (Engelhard, 1996b, 2013). They can be calculated based on the equation below.

$$A_{ij} = \max_{i=1, \dots, I; j=1, \dots, J} \{ |O_{ij} - C_i| \} - |O_{ij} - C_i| \quad (1)$$

where J is the number of raters and I refers to the number of student essays; A_{ij} is the accuracy rating of rater j on essay i ; O_{ij} is the raw rating of Rater j on Essay i ; C_i is the criterion rating on Essay i . Criterion ratings may be defined by a panel of expert raters or the average of observed ratings. In this study, criterion ratings are given by expert raters. Modeling this distance using the accuracy ratings allows us to create a latent continuum of rating accuracy (Engelhard, 1996b).

Unfolding model

In this study, accuracy ratings are modeled using the hyperbolic cosine model (Andrich & Luo, 1993) to define the hyperbolic cosine accuracy model (HCAM) for creating an accuracy continuum (J. Wang et al., 2016). The HCAM can be expressed as below.

$$P(X_{ij} = k) = \frac{[\cosh(\delta_i - \lambda_j)]^{m-k} \prod_{l=1}^k \cosh(\rho_{il})}{\sum_{k=0}^m [\cosh(\delta_i - \lambda_j)]^{m-k} \prod_{l=1}^k \cosh(\rho_{il})} \quad (2)$$

when $k = 0$, $\prod_{l=1}^k \cosh(\rho_{il}) \equiv 1$,

where $k = 0, \dots, m$, and m is the number of categories of accuracy ratings; X_{ij} = observed accuracy rating received by rater j on essay i ; δ_i = difficulty of essay i to score accurately; λ_j = accuracy location of rater j ; ρ_{il} = essay threshold parameter, these threshold parameters are constrained to be equally distanced across accuracy rating categories.

The HCAM explores the difficulty-to-score of essays for individual raters. In other words, raters with different accuracy locations on the unfolding scale may score different subsets of essays accurately. The HCAM achieves this by modeling the relative distance between the essays and raters based on a proximity principle. That said, essays and raters are located along a common unfolding scale, and raters locate closer to the essays that they score more accurately. Figure 1 shows the probability function curves for an essay based on HCAM. The relative location refers to the relative difference between a rater's location and an essay's location. Raters who are located closer to an essay have greater likelihood to score this essay accurately. With the increase in the relative locations, the probability of accurate scoring ($X = 1$) decreases and the probability of inaccurate scoring ($X = 0$) increases. In this study, the essay locations on the unfolding continuum are used for exploring the substantive meaning of the scale with the use of a supervised Latent Dirichlet Allocation model.

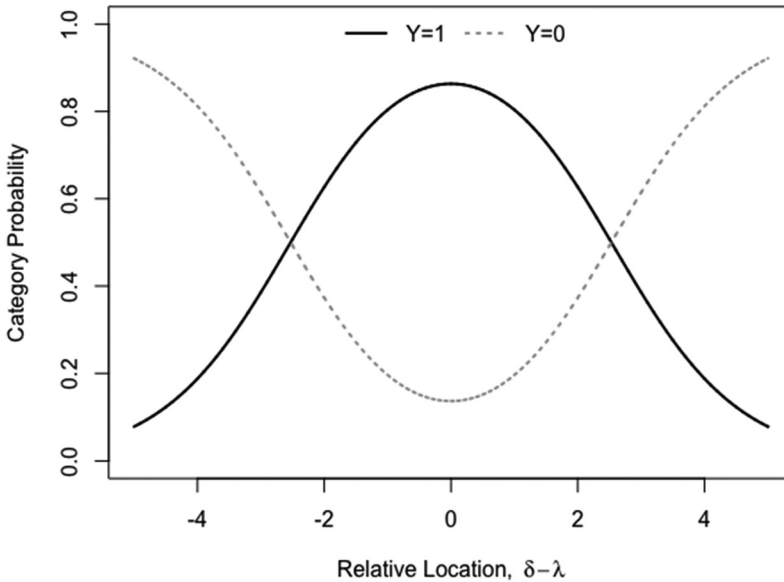


Figure 1. Probability function curves for a hyperbolic cosine accuracy model. $Y = 1$ indicates that there is agreement between operational ratings and expert ratings. $Y = 0$ indicates that there is not agreement between operational ratings and expert ratings.

Supervised latent Dirichlet allocation

Latent Dirichlet Allocation (LDA; D. M. Blei et al., 2003) is a probabilistic mixture model that uses a hierarchical Bayesian design to estimate latent topics within a collection of texts. The hierarchical design works as follows: a collection of text forms the corpus, the corpus consists of a predetermined number of latent topics which are a weighted mixture over a predetermined vocabulary, and each text within the corpus consists of topic proportions which are a weighted mixture over the number of latent topics (D. Blei et al., 2010). For this study, the corpus is formed by the set of essays. The goal of LDA is to estimate the latent topics used throughout the essays and the topic proportions of each essay. The latent topics are clusters of words that are related and represent common components seen across the collection of essays. The topic proportions sum to 1 and represent the usage of each topic. LDA estimates the latent topics and topic proportions through its posterior distribution shown in Equation (3) using a collapsed Gibbs sampling algorithm.

$$P(\boldsymbol{\tau}, \boldsymbol{\beta}, t | w, \eta, \nu) \propto \prod_{n=1}^{N^{(d)}} P(t_{d,n} | \boldsymbol{\tau}_d) P(w_{d,n} | t_{d,n}, \boldsymbol{\beta}_k) \times \prod_{d=1}^D P(\boldsymbol{\tau}_d | \eta) \times \prod_{k=1}^K P(\boldsymbol{\beta}_k | \nu) \quad (3)$$

where $N^{(d)}$ represents the number of words in essay $d \in \{1, 2, \dots, D\}$; $\boldsymbol{\tau}_d$ represents topic proportions for each essay $d \in \{1, 2, \dots, D\}$; $\boldsymbol{\beta}_k$ represents topic $k \in \{1, 2, \dots, K\}$; $t_{d,n}$ represents the topic assignment for each word $n \in \{1, 2, \dots, N^{(d)}\}$ in each essay $d \in \{1, 2, \dots, D\}$; $w_{d,n}$ represents each of the observed words $n \in \{1, 2, \dots, N^{(d)}\}$ in each essay $d \in \{1, 2, \dots, D\}$; η and ν are the prior hyperparameters of the model and represent the concentration of topic proportions in each essay and words within topics, respectively. A small η indicates essays consist of mainly one topic and a large η indicates essays consist of all topics uniformly. A small ν indicates that topics are associated with few words and a large ν indicates that topics are associated with all words uniformly (D. M. Blei & Lafferty, 2009). Since topics are typically unknown beforehand, noninformative prior hyperparameters are often chosen, that is, $\eta = 1$ and $\nu = 1$.

A restriction of LDA is that the estimation of the topics and topic proportions are unsupervised, that is, there are no additional variables that can help drive the inference. However, if we have a dependent variable, such as a score or unfolding location, a supervised latent Dirichlet allocation model (sLDA; McAuliffe & Blei, 2008) can be estimated. The sLDA model uses the dependent variable, that is essay unfolding locations, to help drive inference of the topic assignments, t , from Equation (3). The dependent variable is related to the topic assignments through the following response distribution.

$$y_d | t_{d,1:N^{(d)}}, \mu, \sigma^2 \sim \text{Normal}(\mu' \bar{t}_d, \sigma^2) \quad (4)$$

where y_d represents the dependent variable for essay $d \in \{1, 2, \dots, D\}$ and is assumed to follow a normal distribution with prior means μ and prior variance σ^2 ; \bar{t}_d represents the relative use of each topic in each essay and is defined by $\bar{t}_d = \frac{1}{N^{(d)}} \sum_{n=1}^{N^{(d)}} t_{d,n}$.

The sLDA models the relationship between the topic proportions of each essay and the unfolding measures through a regression model. Results can indicate the degree to which each topic is associated with the essay unfolding measures.

Data source

The essays were written responses to a seventh grade informational extended-response item from an English Language Arts (ELA) formative writing assessment. The extended-response item provided two passages, Passage A and Passage B, and asked the students to write an informative essay that answers the prompt by citing two examples from each passage as evidence. Please see the Appendix for Passages A and B.

Prior to this study, each essay was given an operational score by well-trained operational raters using a rubric-based scoring system. A sample of 100 essays were selected and given a criterion score by a group of expert raters. The distance between the operational scores and criterion scores were used to calculate the accuracy rating of each essay based on Equation (1), and the dichotomous accuracy ratings were analyzed with the HCAM (1 = accurate, 0 = inaccurate).

The sample of 100 essays were then used to fit a sLDA model using the *lda* R package (Chang & Chang, 2010). Prior to fitting the sLDA model, the essays were put through a data pipeline that performed numerous data cleaning tasks. First, all words were changed to lowercase and the punctuation was removed. Next, all stop words were removed from each essay. Stop words are common words used throughout the essays and carry little information about the topics being used (Wilbur & Sirotkin, 1992). For example, common stop words seen in essays written by students include *a*, *are*, *can*, *so*, *the*, *will*, and *you*. Stop words, if not removed, tend to dominate the estimated topics and reduce the interpretability of the model (Choi et al., 2017). Finally, all words were stemmed and corrected for spelling. Stemming retrieves the base word from the different variations of the same word due to the different tenses or grammatical number. The stemming process increases the clarity of the topics and the interpretability of the model (Schofield et al., 2017). These data cleaning tasks provide clearer results and are essential for fitting topic models.

The sLDA model requires the number of latent topics to be determined a priori. Selecting an appropriate number of topics is nontrivial. Common selection techniques for the number of topics are perplexity or the accuracy of a downstream task (D. M. Blei et al., 2003), the deviance information criterion (DIC; Spiegelhalter et al., 2002), and the Watanabe-Akaike information criterion (WAIC; Watanabe & Opper, 2010). Wheeler et al. (2020) used a simulation study to show the appropriate number of topics with differing amounts of data responses. Given the number of essays and essay lengths of our sample, a three-topic model could be estimated and provide stable measures. Thus, following recommendations from Wheeler et al. (2020), along with comparing DIC values, a three-topic sLDA model was chosen for the analysis in this study.

Analysis plan

The accuracy ratings were used to create a latent continuum with the HCAM for defining rater accuracy and difficulty-to-score of essays. The set of essays were analyzed with a sLDA model to estimate the latent topics and their relationship to the latent continuum defined by the HCAM. In order to better understand the latent topics, a qualitative analysis of the essays was performed by reading individual essays that primarily used one of the three topics. It is important to note that latent topics are similar to latent factors in factor analysis, and that the term latent topics is not directly related to traditional definitions of topics in the writing assessment literature. Additionally, the source articles – Passage A and Passage B provided by the extended-response item were analyzed with the fitted three-topic sLDA model to assist in defining the latent topics. Finally, the topic proportions for each essay were used in a multivariate regression with the essays' unfolding locations as the dependent variable to determine which semantic features could explain why essays are more or less difficult to score accurately.

Results

The primary purposes of this study are to identify the latent topics used by students when responding to this particular extended-response item and to identify which of these topics explain the difficult-to-score essays on the unfolding continuum.

Unfolding groups

The 100 essays are centered at zero on the unfolding scale with a standard deviation of 1.72. The essay locations range from -4.30 to 3.50 . There is a polynomial relationship between accuracy rates and

unfolding locations for essays. A higher accuracy rate shows that an essay was easier for raters to score accurately. Therefore, a closer-to-zero unfolding location measure indicates easier-to-score, and a more extreme unfolding location measure implies more difficult-to-score (Figure 2).

The unfolding continuum differentiated difficult-to-score essays into two directions, and we named them as difficult-to-score below and difficult-to-score above. We further separated the essays into three groups by their unfolding locations (Figure 3). It is somewhat easier to see the effects related to the topics when the essays are categorized with respect to the unfolding continuum. The difficult-to-score below essays ($n = 24$) were defined as having an unfolding location less than -1.5 . Essays in the difficult-to-score below group had accuracy rates between 0.45 to 0.80, indicating that there was discrepancy between the scores given by the operational raters and the scores given by expert raters. The easy-to-score essays ($n = 50$) had an unfolding location between -1.5 and 1.5 . Essays in the easy-to-score groups had accuracy rates between 0.80 and 0.98, indicating that there was little discrepancy between operational and expert ratings. The difficult-to-score above essays ($n = 26$) were categorized as having an unfolding location greater than 1.5 . Similar to the difficult-to-score below group, essays in the difficult-to-score above group had accuracy rates between 0.50 and 0.80, indicating discrepancy between operation and expert ratings.

Latent topics and topic proportions

The latent topics help define the unfolding continuum, and explore what students are writing about that causes their essays to be difficult to score differently. Particularly, the sLDA models each topic and estimates the probability of each word occurring. By looking at the 25 words with the highest probability for each topic, we get a better sense of the substantive meaning of each topic. Table 1 shows the top 25 words for each of the three topics estimated in the sLDA model with estimated probability of that word occurring under the given topic. It is important to note that although the probabilities of the word occurring look small, they are relatively large. For instance, since the total number of unique words across all essays is 421, if all words under a topic were equally likely to occur, the probability would be $\frac{1}{421} = 0.002$. This means a word with a probability of 0.012 is 6 times more

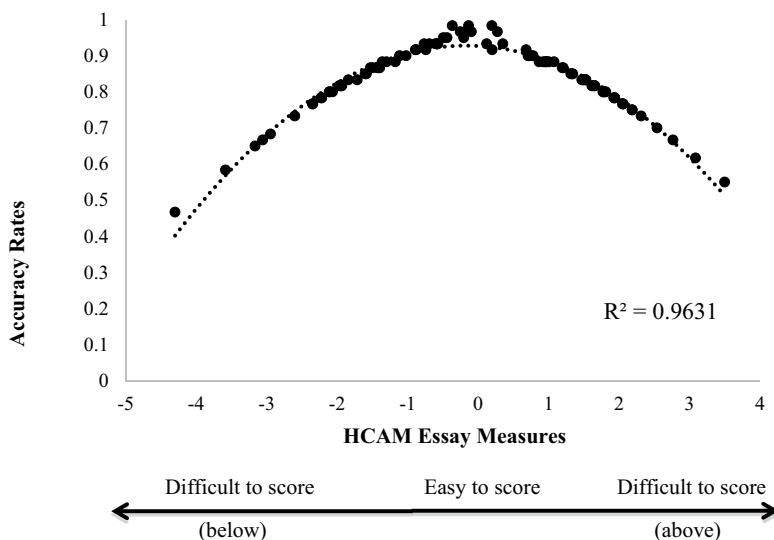


Figure 2. Relationship between hyperbolic cosine accuracy model (HCAM) essay measures and observed accuracy rates. The accuracy rate is calculated using the sum of accuracy ratings for an essay by all raters divided by the maximum possible points. A second-order polynomial curve explains 96.31% variation in this relationship.

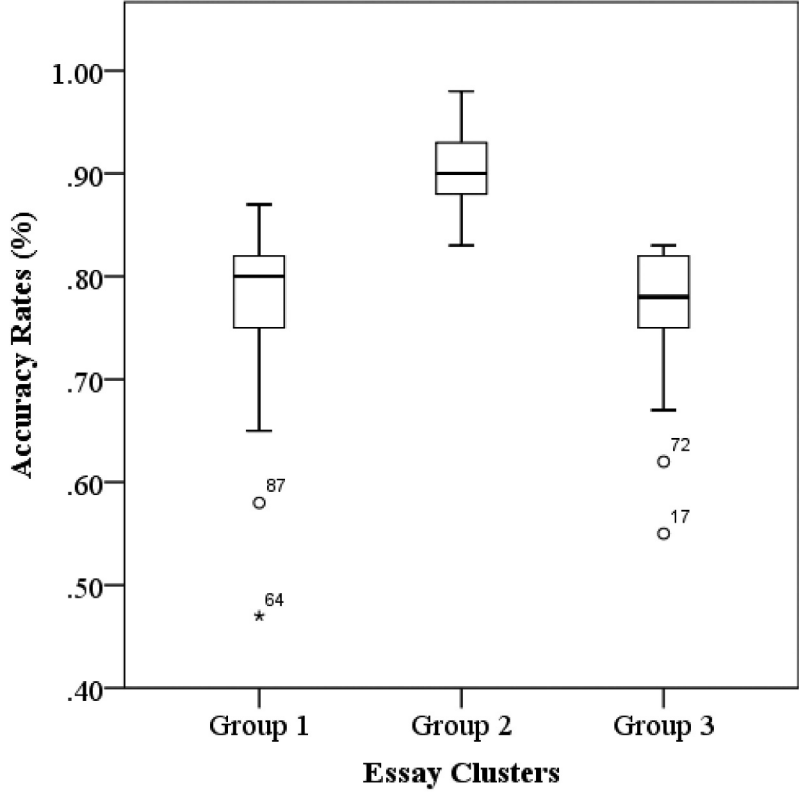


Figure 3. Distribution of accuracy rates within each essay cluster. Group 1 – difficult-to-score below, Group 2 – easy-to-score, and Group 3 – difficult-to-score above.

likely to occur than random. Therefore, [Table 1](#) also provides a probability ratio to express how much more likely a word is to occur compared to a random selection from a uniform distribution.

The three topics estimated from the sLDA model have quite a bit of overlap, for example, the word *water* appears at the top of each topic. This is expected with extended-response items since the answers are constrained by the prompt (Choi et al., 2017). The sLDA model, however, is able to pick up on the subtle nuances of how these words are being used by the student and is able to distinguish between the topics. By doing so, the model provides topic proportions for each essay, which represents the usage of each topic in each essay. The sLDA model estimated that the average topic proportion across all essays for Topic One was 0.289(0.287), the average topic proportion for Topic Two was 0.474(0.331), and the average topic proportion for Topic Three was 0.237(0.242). From this result, we see that overall Topic Two was used the most, followed by Topic One, and then Topic Three.

The analysis of topic proportions can be further broken down into the unfolding groups ([Figure 4](#)). From this view, we see that the difficult-to-score below group (Group 1) has a high average proportion of Topic One and a low average proportion of Topic Three. This means that the essays from the difficult-to-score below group mainly used words of Topic One. Similarly, the easy-to-score group (Group 2) and the difficult-to-score above group (Group 3) have a high average Topic Two and Topic Three proportion, respectively. This means that the easy-to-score group is mainly using words of Topic Two while the essays of the difficult-to-score above group is mainly using words in Topic Three.

Table 1. The top 25 words and their probabilities of occurring for Topic 1, Topic 2, and Topic 3.

Topic 1			Topic 2			Topic 3		
Word	Probability	Probability Ratio	Word	Probability	Probability Ratio	Word	Probability	Probability Ratio
1 Water	0.110	46.31	Water	0.097	40.84	Water	0.104	43.78
2 Africa	0.029	12.21	School	0.028	11.79	Oil	0.030	12.63
3 Many	0.026	10.95	Children	0.025	10.53	Africa	0.020	8.42
4 People	0.024	10.10	Africa	0.024	10.10	People	0.019	8.00
5 Oil	0.021	8.84	Country	0.021	8.84	Crisis	0.014	5.89
6 Drink	0.019	8.00	Because	0.021	8.84	Supply	0.014	5.89
7 Pollute	0.015	6.32	Ease	0.019	8.00	Contaminate	0.014	5.89
8 Because	0.014	5.89	Time	0.017	7.16	More	0.012	5.05
9 Crisis	0.012	5.05	Unclean	0.016	6.74	Die	0.012	5.05
10 Waste	0.012	5.05	Crisis	0.014	5.89	Cause	0.011	4.63
11 Children	0.012	5.05	People	0.014	5.89	Spill	0.011	4.63
12 Contaminate	0.012	5.05	State	0.013	5.47	Was	0.010	4.21
13 Country	0.012	5.05	Drink	0.012	5.05	Passage	0.010	4.21
14 These	0.012	5.05	cause	0.012	5.05	Most	0.009	3.79
15 Make	0.010	4.21	Clean	0.011	4.63	Paragraph	0.009	3.79
16 Animal	0.009	3.79	Consequence	0.011	4.63	Hole	0.009	3.79
17 Need	0.009	3.79	Many	0.011	4.63	Walk	0.008	3.37
18 Sick	0.009	3.79	Her	0.010	4.21	Etana	0.008	3.37
19 Clean	0.009	3.79	Etana	0.010	4.21	Reason	0.008	3.37
20 Find	0.009	3.79	Passage b	0.010	4.21	Pollute	0.008	3.37
21 Consequence	0.008	3.37	More	0.009	3.79	Passage A	0.008	3.37
22 Mile	0.008	3.37	Passage a	0.009	3.79	Effect	0.008	3.37
23 Day	0.008	3.37	Miss	0.009	3.79	Bug	0.007	2.95
24 Dump	0.008	3.37	These	0.009	3.79	Day	0.007	2.95
25 Kill	0.008	3.37	Day	0.009	3.79	Many	0.007	2.95

A uniform distribution over words would Carry a probability of 0.002, in which case majority of the top 25 words are more likely to occur by a factor larger than 4. The *Probability Ratio* column indicates how much more likely each word is to occur compared to a random uniform distribution.

sLDA regression results

Along with the estimated topics and topic proportions, the sLDA model captured the relationship between the unfolding locations and the topics through a regression analysis. Each essay was considered as an observation; thus, our regression analysis had 100 observations ($n = 100$). The unfolding location for each essay was the dependent variable and ranged between -4.3 and 3.5 .

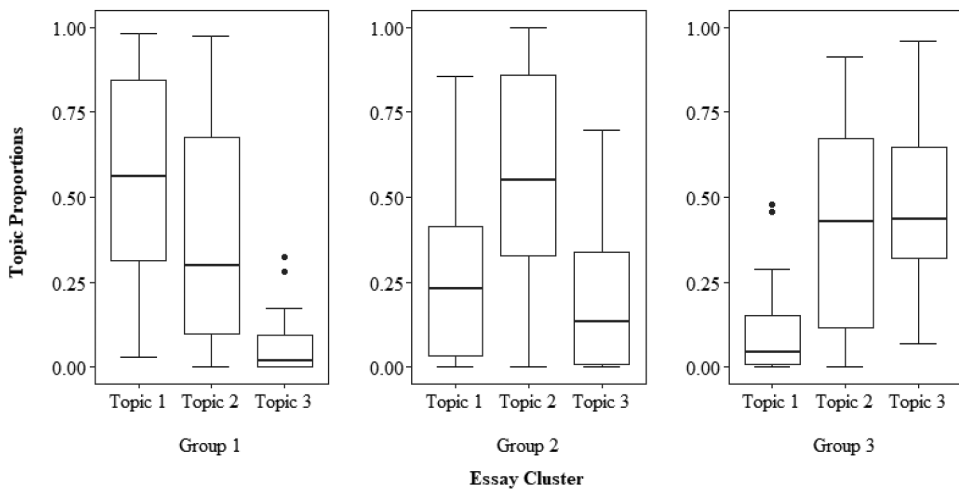


Figure 4. Box plots for the usage of each topic with respect to the unfolding groups. The sum of the topic proportions for each line equals 1. Group 1 – difficult-to-score below, Group 2 – easy-to-score, and Group 3 – difficult-to-score above.

Table 2. sLDA regression analysis: topic proportion effects on unfolding location.

Effect	Estimate	SE	95% CI		<i>p</i>
			<i>LL</i>	<i>UL</i>	
Topic One	−2.712	0.313	−3.339	−2.087	<.001
Topic Two	−0.143	0.218	−0.579	0.293	0.515
Topic Three	3.593	0.385	2.823	4.363	<.001

Independent variables were the topic proportions of all three topics for each essay. Dependent variable was the unfolding location measures. Topic proportion values ranged from 0 to 1. Unfolding location values ranged from −4 to 4. CI = confidence interval; *LL* = lower limit; *UL* = upper limit.

The topic proportions for the three topics for each essay were our independent variables and ranged between 0 and 1. Table 2 provides the sLDA regression results. The regression model showed that the topic proportions could moderately explain the variation in the unfolding location ($r^2 = 0.561$) and found that Topics One and Three had a significant linear relationship with unfolding location measures ($p < 0.001$) while Topic Two did not have a significant linear effect ($p = 0.515$). Furthermore, the regression coefficient for Topic One was negative ($\beta_{Topic1} = -2.712$), indicating that the difficult-to-score below group have higher proportions on Topic One. The regression coefficient for Topic Three was positive ($\beta_{Topic3} = 3.593$), indicating that the difficult-to-score above group have higher proportions on Topic Three. Although Topic Two did not show a significant linear effect on the unfolding measures, it may have other types of relationship (e.g., polynomial) with the unfolding continuum.

The relationship between the topics and the unfolding groups is easier to see through box plots shown in Figure 4. That is, Topic One was mainly used by the difficult-to-score below group, Topic Two was mainly used by the easy-to-score group, and Topic Three was mainly used by the difficult-to-score above group.

Defining latent topics

The sLDA model was able to distinguish between the topics in each response. In order to further define the topics, a qualitative analysis of each essay was performed. Specifically, we read essays from the three unfolding groups and analyzed the unfolding location, topic proportions, and raw text. Table 3 provides three responses from the 100 essays in this study. Student A's essay belongs to the difficult-to-score below group and primarily used Topic One in their response. This essay response did use the evidence from the passages provided, however, it failed to provide specific citations to the corresponding passages. Student B's essay belongs to the easy-to-score group and primarily used Topic Two in the response. This essay response followed the instruction by using and citing the evidence from both source articles to support its own argument. Student C's essay belongs to the difficult-to-score above group and primarily used Topic Three in their response. This essay response cited the examples from one of the passages provided, namely, Passage A.

To further demonstrate the connection between the textual borrowing feature of an essay and how difficult it is for raters to score accurately, we fit the sLDA model to the source passages provided in the extended-response item. The appendix provides both passages and their estimated topic proportions. The results show that Passage A mainly uses Topic Two ($\tau_{PassageA, Topic2} = 0.536$) and Topic Three ($\tau_{PassageA, Topic3} = 0.389$) while Passage B mainly uses Topic Two ($\tau_{PassageB, Topic2} = 0.874$). In other words, Topic Two is related to both of the passages, Topic Three is related to Passage A only, and Topic One is not necessarily related to either of the passages.

The unfolding continuum differentiated the essays into two directions that are difficult-to-score below and difficult-to-score above. The essays with different unfolding locations used different latent topics in the responses. Table 4 shows a meaningful name and description to define each of the three topics. The extended-response item asks the students to use both the passages provided and use two pieces of evidence from each passage. Our results suggest that Topic One is for students who used evidence in the

Table 3. Sample responses with unfolding location and topic proportions.**Student A's Essay – Difficult-to-Score Below Group (Unfolding Location = -1.969)****Topic 1 Proportion** = 0.842, **Topic 2 Proportion** = 0.140, **Topic 3 Proportion** = 0.018

In Africa the water is very polluted and dangaroose. Many Africans die every day from illnesses and mostly dehydration. Since all of the fresh water of polluted rely nothing is safe to drink in africa. So any water they do drink its unsafe. When they get sick they have to miss school for a certain amount of time. They stay out until there healthy again. Wich could take a long time. The illneses are very deadly if not if not treated right. Their working on ways to convert a solid into a liqid. A human being can only last 3–4 days without fresh, clean water. Not only that drought occures very often in africa. Sometimes all you can find is a mud hole in the bottom of a dried up water bed. That all the info I could find.

Student B's Essay – Easy-to-Score Group (Unfolding Location = 0.966)**Topic 1 Proportion** = 0.067, **Topic 2 Proportion** = 0.933, **Topic 3 Proportion** = 0

... The first conqense I will be talking about is children not being able to go to school. "Because she has to make this trip two more times today, there is no time to go to school." This quote from **passage B** is from when Etona is fetiching water from a water hole. It begins to say she has no time to go to school, which means she won't get into college; that means she will not get a job (or at least one that pays much); that will hurt the economy. A statement from **passage A** that further supports that which I am saying is "children in some African countries often miss school because of illnesses resulting from unclean water and poor sanitation practices." This further support my statement saying that the water that taking long periods of time to fetch can make you sick causing you to miss more days of school. This brings me into my second consequence.

Secondly, people become sick from unclean water. My reasoning for this is this statement from **passage A**. "More than 85% of all dieseses in African children are caused by unclean water." This states that nearly all of dieseses children can get are result of unclean water. This will lead to a much higher mortality rate as well. I can support this with these statements from **passage B**. "She has known many who have become sick from drinking polluted water. Some of her friends have died from dieseses they contracted from the polluted water." This statement (as previously said) supports my idea at large amonts of sicknesses and deaths from those sicknesses in Africa ...

Student C's Essay – Difficult-to-Score Above Group (Unfolding Location = 2.314)**Topic 1 Proportion** = 0.153, **Topic 2 Proportion** = 0.106, **Topic 3 Proportion** = 0.741

... In some parts of Africa, the urbanization and oil spills are some factors that happen to contribute to the water crisis. "As more and more people move closer to populated cities, the food supply demand grows. Farming and agriculture have increased, creating the need for more fertilizers that are damaging the water supply." This piece of textual evidence from **Passage A** is one cause for the countries in Africa to have low water supply. Sadly, that is not the only cause. "Nigeria is Africa's largest oil producer. Last year, 6,000 tons of oil were dumped into the Niger Delta waterway." This textual evidence from **Passage A** is another cause to add to the list. Since the oil just sits on top of the water, it kills and contaminates alot of the plants and animals. The animals don't really have a choice to drinking it or not because there is not much water source around them, so they have to drink what they have. One more consequence of the water crisis is that a person can only live for so long without water. The estimated amount of days are 3–4. Another consequence is that alot of Africans have to walk several miles each day in order to get the water to support their family. "Her village has no running water, so Etana and the other women and children must make a daily trek to get water." This shows that many people have to walk to get water. "She is on her way to the only water source she knows, which is located several miles away." This shows that they had to walk several miles ...

One sample essay from each of the unfolding location groups. All passage references are bolded.

Table 4. Topic names and descriptions.

Topic	Name	Description
Topic One	Provided Evidence without Citing Source	Students who used Topic One provided evidence but did not follow the prompt by not citing either of the passages provided
Topic Two	Provided Evidence with Citing Source (Passage A and Passage B)	Students who used Topic Two provided sufficient evidence and followed the prompt by citing both passages provided
Topic Three	Provided Evidence with Citing Source (Passage A)	Students who used Topic Three provided evidence but did not fully follow the prompt by citing only one of the passages provided

passages but failed to cite their sources. Failing to cite the sources caused the essays to be difficult-to-score accurately because the student answered the prompt but did not completely follow instructions on composition. Topic Two is for students who used evidence in both passages and cited their sources. Since the students answered the prompt and followed the instructions, their essays were easy-to-score accurately. Topic Three is for students who used evidence from Passage A and cited their source, but failed to use both passages. Failing to use both passages caused essays to be difficult-to-score accurately because the student answered the prompt and provided evidence but did not completely follow instructions.

Discussion

This study proposed the use of a sLDA model to empirically define the substantive meaning of an accuracy unfolding continuum. Student essays and raters were placed onto the common unfolding scale. The location measures were obtained based on the hyperbolic cosine accuracy model with the use of accuracy ratings. Results based on the sLDA regression analysis and box plots showed that all three of the topics have a moderate influence on the unfolding location measures. Specifically, Topic One is shown to be associated with essays that are difficult-to-score below, that is, student essays that mainly consisted of words in Topic One were difficult to score accurately. Topic Two is shown to be associated with essays that are easy-to-score, that is, the essays that used more of Topic Two were easy to score accurately. Topic Three is shown to be associated with essays that are difficult-to-score above, that is, these essays were also difficult to score accurately.

A qualitative analysis of the essays and the passages provided from the extended-response item further defined the latent topics. The easy-to-score essays generally had highest topic proportion on Topic Two, and the responses strictly followed the instruction that (a) cite at least two examples from Passage A, (b) cite at least two examples from Passage B, and (c) explain how these examples illustrate the discussed topic (i.e., consequences of water crisis). The essays of the difficult-to-score below group typically had highest topic proportion on Topic One. These essay responses used the examples but failed to cite the passages provided. On the other hand, the essays of the difficult-to-score above group had highest topic proportion on Topic Three, and the responses cited examples from mainly Passage A.


It is worth noting that easy-to-score essays are those essays that raters possess more consistent and accurate judgments about the reflected level of writing proficiency. On the contrary, when raters possess more inconsistent and inaccurate judgments toward an essay, the essay becomes more difficult-to-score. Based on the accuracy ratings, the unfolding accuracy continuum shows how difficult for raters to score an essay accurately, rather than the level of writing proficiency of each student.

The results from the topic models indicate interpretation for this particular writing assessment. That is, the estimated sLDA model is dependent on the prompt. This is important to note since it limits the generalizability of the results. The analysis, however, suggested one potential, generalizable reason why an essay is either difficult-to-score accurately or easy-to-score accurately based on the content topics. J. Wang et al. (2017) explored rater perceptions during the scoring of an integrated writing assessment and found that raters did not have the same understanding of the appropriate amount of textual borrowing and ways of citing the evidence from source articles. This supports our findings of the relationship between latent topics and the unfolding accuracy continuum. The latent topics reflect different amount of textual borrowing as well as the citing sources. Inconsistent perception among raters toward this essay feature is identified as a source of inaccuracy.

This study provides an approach based on unfolding models and sLDA models for exploring different reasons that some essays are more difficult to score accurately than others. Our findings can help researchers, educators, and classroom teachers in writing assessments better understand the relationship between essay characteristics and rater scoring accuracy, which will further identify sources of inaccuracy and improve the rating quality.

ORCID

Jordan M. Wheeler  <http://orcid.org/0000-0001-9766-5014>

George Engelhard  <http://orcid.org/0000-0002-1694-8942>

Jue Wang  <http://orcid.org/0000-0002-3519-2693>

References

- Andrich, D., & Luo, G. (1993). A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses. *Applied Psychological Measurement*, 17(3), 253–276. <https://doi.org/10.1177/014662169301700307>

- Aubin, A. S., St-Onge, C., & Renaud, J. S. (2018). Detecting rater bias using a person-fit statistic: A Monte Carlo simulation study. *Perspectives on Medical Education*, 7(2), 83–92. <https://doi.org/10.1007/s40037-017-0391-8>
- Blei, D., Carin, L., & Dunson, D. (2010). Probabilistic topic models. *IEEE Signal Processing Magazine*, 27(6), 55–65. <https://doi.org/10.1109/MSP.2010.938079>
- Blei, D., & Lafferty, J. (2009). Topic models. In A. N. Srivastava & M. Sahami (Eds.), *Text mining: Classification, clustering, and applications* (pp. 101–124). CRC press.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan), 993–1022. <https://dl.acm.org/doi/pdf/10.5555/944919.944937>
- Cardozo-Gaibisso, L., Kim, S., Buxton, C., & Cohen, A. (2020). Thinking beyond the score: Multidimensional analysis of student performance to inform the next generation of science assessments. *Journal of Research in Science Teaching*, 57(6), 856–878. <https://doi.org/10.1002/tea.21611>
- Chang, J., & Chang, M. J. (2010). lda: Collapsed Gibbs sampling methods for topic models. R package version 1.4.2. <https://CRAN.R-project.org/package=lda>
- Choi, H. J., Kwak, M., Kim, S., Xiong, J., Cohen, A. S., & Bottge, B. A. (2017, July). An application of a topic model to two educational assessments. In *The Annual Meeting of the Psychometric Society* (pp. 449–459). Springer.
- Duong, E., Mellom, P., & Hixon, R. (2019). Using topic modeling to analyze the effects of instructional conversation on 3rd grade students' writing. Paper presented at the annual meeting of the American Association for Applied Linguistics, Atlanta, GA.
- Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly: An International Journal*, 2(3), 197–221. https://doi.org/10.1207/s15434311laq0203_2
- Engelhard, G., Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, 31(2), 93–112. <https://doi.org/10.1111/j.1745-3984.1994.tb00436.x>
- Engelhard, G. (1996a). Constructing rater and task banks for performance assessments. *Journal of Outcome Measurement*, 1(1), 19–33. <https://europaemc.org/article/med/9661713>
- Engelhard, G. (1996b). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement*, 33(1), 56–70. <https://doi.org/10.1111/j.1745-3984.1996.tb00479.x>
- Engelhard, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Routledge.
- Kim, S., Kwak, M., Cardozo-Gaibisso, L., Buxton, C., & Cohen, A. S. (2017). Statistical and qualitative analyses of students' answers to a constructed response test of science inquiry knowledge. *Journal of Writing Analytics*, 1(1). <https://doi.org/10.37514/JWA-J.2017.1.1.05>
- Mcauliffe, J. D., & Blei, D. M. (2008). Supervised topic models. In J. C. Platt, D. Koller, Y. Singer, & S. T. Roweis (Eds.), *NIPS 2007, Advances in neural information processing systems* (pp. 121–128). Curran Associates.
- Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, 4(4), 386–422. PMID: 15064538.
- Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics*, 27(4), 341–384. <https://doi.org/10.3102/10769986027004341>
- Read, B., Francis, B., & Robson, J. (2005). Gender, 'bias', assessment and feedback: Analyzing the written assessment of undergraduate history essays. *Assessment & Evaluation in Higher Education*, 30(3), 241–260. <https://doi.org/10.1080/02602930500063827>
- Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing*, 15(1), 18–39. <https://doi.org/10.1016/j.asw.2010.01.003>
- Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, 88(2), 413. <https://doi.org/10.1037/0033-2909.88.2.413>
- Schofield, A., Magnusson, M., & Mimno, D. (2017). Understanding text pre-processing for latent Dirichlet allocation. In *Proceedings of the 15th conference of the European chapter of the Association for Computational Linguistics* (Vol. 2, pp. 432–436). Valencia, Spain.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 64(4), 583–639. <https://doi.org/10.1111/1467-9868.00353>
- Wang, J., Engelhard, G., Jr, Raczynski, K., Song, T., & Wolfe, E. W. (2017). Evaluating rater accuracy and perception for integrated writing assessments using a mixed-methods approach. *Assessing Writing*, 33, 36–47. <https://doi.org/10.1016/j.asw.2017.03.003>
- Wang, J., Engelhard, G., Jr, & Wolfe, E. W. (2016). Evaluating rater accuracy in rater-mediated assessments using an unfolding model. *Educational and Psychological Measurement*, 76(6), 1005–1025. <https://doi.org/10.1177/0013164415621606>
- Wang, J., & Engelhard, J. G. (2019a). Conceptualizing rater judgments and rating processes for rater-mediated assessments. *Journal of Educational Measurement*, 56(3), 582–609. <https://doi.org/10.1111/jedm.12226>

- Wang, J., & Engelhard, J. G. (2019b). Exploring the impersonal judgments and personal preferences of raters in rater-mediated assessments with unfolding models. *Educational and Psychological Measurement*, 79(4), 773–795. <https://doi.org/10.1177/0013164419827345>
- Wang, W. C., Su, C. M., & Qiu, X. L. (2014). Item response models for local dependence among multiple ratings. *Journal of Educational Measurement*, 51(3), 260–280. <https://doi.org/10.1111/jedm.12045>
- Watanabe, S., & Oppen, M. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11, 12. <https://www.jmlr.org/papers/volume11/watanabe10a/watanabe10a.pdf>
- Wheeler, J. M., Cohen, A. S., Xiong, J., Lee, J., & Choi, H.-J. (2020). A simulation guide for topic models of constructed-response items [Paper presentation]. (Virtual) International Meeting of the Psychometric Society, College Park, MD.
- Wilbur, W. J., & Sirotkin, K. (1992). The automatic identification of stop words. *Journal of Information Science*, 18(1), 45–55. <https://doi.org/10.1177/016555159201800106>
- Wilson, M., & Hoskens, M. (2001). The rater bundle model. *Journal of Educational and Behavioral Statistics*, 26(3), 283–306. <https://doi.org/10.3102/10769986026003283>
- Wind, S. A., & Engelhard, G. (2012). Examining rating quality in writing assessment: Rater agreement, error, and accuracy. *Journal of Applied Measurement*, 13(4), 321–335. PMID: 23270978.
- Wolfe, E. W., & McVay, A. (2012). Application of latent trait models to identifying substantively interesting raters. *Educational Measurement: Issues and Practice*, 31(3), 31–37. <https://doi.org/10.1111/j.1745-3992.2012.00241.x>

Appendix

Passage A: concerns about water supply in parts of Africa

(Topic 1 Proportion = 0.075, Topic 2 Proportion = 0.536, Topic 3 Proportion = 0.389)

Lack of clean water is one of the most critical challenges facing some African countries. Limited freshwater also leads to environmental issues that are harmful to the world's ecological balance. Less than 10% of Africa's surface water is available for its citizens. Drought conditions make surface water scarce. The majority of the water is under large rock formations. Most water in remote regions is accessed from open holes dug in the sand of dry riverbeds. Due to this, some citizens walk many miles to find clean drinking water. Children in some African counties often miss school because of illnesses resulting from unclean water and poor sanitation practices. More than 85% of all diseases in African children are caused by using unclean water. The most common diseases transmitted through its water sources are Hepatitis and Typhoid Fever. In exchange for money, some African countries accept solid waste from developed countries in Asia, Europe, and North America. However, the governments of these African countries have generally not developed adequate solid waste treatment. The waste is often dumped into rivers and bodies of water. Other Causes of the Water Crisis Urbanization and oil spills are also factors that contribute to the water crisis in parts of Africa. As more and more people move closer to populated cities, the food supply demand grows. Farming and agriculture have increased, creating the need for more fertilizers that are damaging the water supply. Furthermore, oil spills can be devastating. Nigeria is Africa's largest oil producer. Last year, 6,000 tons of oil were dumped into the Niger Delta waterway. The oil floats on the waterways, killing and contaminating plants and animals. Oil has been released into the ocean by damaged pipelines. The coastlines by the fishing villages are slick with oil. The people then drink or bathe in the contaminated water. A person can only live 3–4 days without clean water. Making water safer is a key challenge facing some African countries.

Passage B: The Story of Etana

(Topic 1 Proportion = 0.048, Topic 2 Proportion = 0.874, Topic 3 Proportion = 0.078)

There is no time for school again today! Etana wakes up very early, just before sunrise. She heads out the door like she does three times a day. Her village has no running water, so Etana and the other women and children must make a daily trek to get water. Etana grabs her “jerry can” and hoists it above her head. She is on her way to the only water source she knows, which is located several miles away. The road is hot, and the temperature is hot. It takes her almost an hour to make the trip each time. The water source is a small muddy hole in a riverbed. The water hole is filled with debris, bugs, and bacteria. Etana has no choice but to fill her can. If she doesn't stop here, there is no guarantee that she will be able to find another water supply. She notices a cow standing nearby, waiting for his turn to drink from the watering hole. After filling her can with water, it becomes extremely heavy; she has to stop and rest many times. The jerry can may weigh as much as 40 pounds when it is filled with water. She has known many people who have become sick from drinking the polluted water. Some of her friends have died from diseases they contracted from the polluted water. After Etana brings water home, it must be boiled before it can be used for cooking, bathing, or drinking. Because she has to make this trip two more times today, there is no time to go to school. Etana wants to be a doctor when she grows up. This is important to Etana because she wants to help the people in her community. She aspires to learn about ways to help the people make the water safer to use. Her dream of being a doctor will have to wait; maybe tomorrow she will be able to go back to school.