Operational Automatic Scoring of Text Responses in 2016 ePIRLS: Performance and Linguistic Variance

Hyo Jeong Shin[1], Nico Andersen[2], Andrea Horbach[3,4], Euigyum Kim[5], Jisoo Baik[1], and Fabian Zehner[2,6]

[1]Sogang University, South Korea

[2]DIPF | Leibniz Institute for Research and Information in Education, Germany

[3]FernUniversität Hagen, CATALPA, Germany

[4]University of Hildesheim, Germany

[5]Upstage, South Korea

[6]Centre for International Student Assessment (ZIB), Germany

Abstract

In this project report, we report on the feasibility of automatic scoring systems for text responses from the 2016 ePIRLS. We show that the multilingual automatic scoring approach used in this study can be applied to different languages and countries, despite their linguistic variance. To measure linguistic variance, we used a variant of the conventional type-token-ratio, which we refer to as STTR. We utilized two systems for automatic scoring: fuzzy lexical matching (FLM) and supervised classifiers based on semantics. FLM prioritizes accuracy but requires significant manual scoring work by human raters. The supervised classifiers were trained using a pre-trained deep neural network (XLM-R) for multilingual texts and support vector machines. Results showed that automatic scoring models can score accurately ($\kappa = .755$ on average using XLM-R) and efficiently (26.1% reduction of manual scoring on average) across languages and countries, in the presence of linguistic variance. However, performance varied widely across items, highlighting the importance of investigating the determinants of automatic scoring performance. It was found that higher levels of linguistic variance were associated with lower automatic scoring performance. In addition, linguistic variance and automatic scoring model performance were significantly related to several item- and student-level characteristics. The paper concludes with a discussion of the implications of operationalizing automatic scoring.

Introduction

Educational assessments commonly comprise a substantial number of constructed-response (CR) items that require test takers to compose short text responses. These responses are then evaluated with respect to predefined scoring guides. With the rise of nascent technologies in natural language processing, researchers and assessment providers are looking more and more into the feasibility of assisting or even replacing their human scorers by means of automatic scoring (e.g., Yamamoto, He, Shin, & von Davier, 2018; Yaneva & von Davier, 2023; Whitmer et al., 2023; Zehner, Sälzer, & Goldhammer, 2016; Sukkarieh, Von Davier, & Yamamoto, 2012). Relatedly, after PISA switched to CBA in 2015, PISA 2018 operationalized a machine-supported coding system (MSCS) that allowed for the automatic coding of approximately 25% of text responses across all countries, languages, and domains (OECD, 2019). With the motivation to be generalizable across multiple languages, the MSCS utilizes exact-matching and automatically applies verified labels to incoming unscored responses. A study by Zehner et al. (2021) revealed that automatic normalization steps could score more text responses (+5.1%) with only a minor loss (-0.5%) in accuracy compared to the MSCS.

For international large-scale assessments such as PIRLS and PISA, this feasibility of operational automatic scoring remains challenging; among others, because of the assessments' massively multilingual nature, the varying performance of automatic scoring across items, and the assessments' high stakes at the policy level. Besides factors determined by the setting (Zesch, Horbach, & Zehner, 2023), the variability in building accurate classifiers across items is often attributed to the corresponding differences in linguistic variance in the text responses elicited by each item (Horbach & Zesch, 2019). That is, the more constrained the responses are linguistically, the easier it is for the scoring model to pick up the patterns that determine a response's correctness from the training data and apply accurate classifications to new data. Consider, for example, the limited variation in responses to a fictional item such as *"Why do people usually see the lightning before hearing the thunder?"* opposed to the large universe of responses to an item such as *"What is the message of the story?"*. We distinguish three theoretical components of linguistic variance in short text responses (Zesch et al., 2023): conceptual variance (i.e., the concepts elicited by an item), realization variance (i.e., the linguistic expression of semantic concepts and their relation; including paraphrases, misspellings, etc.), and nonconformity variance (i.e., aberrant responses).

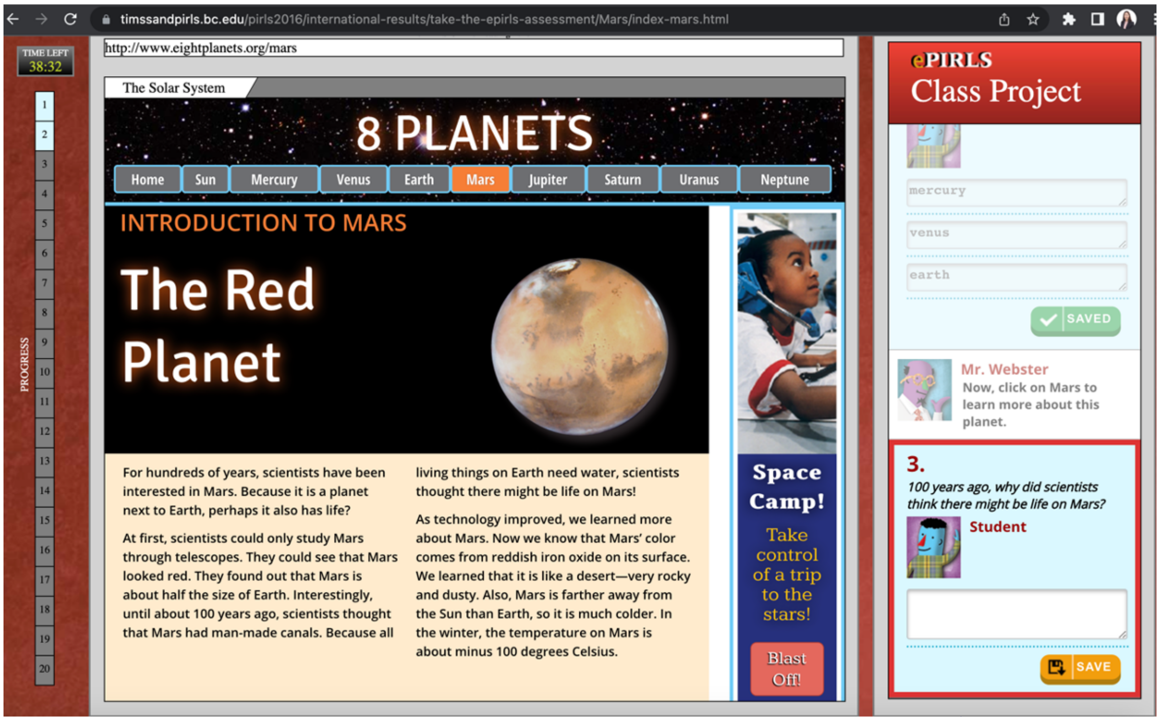
Therefore, this project sets out (1) to apply automatic scoring to text responses from the 2016 ePIRLS to examine the performance of automatic scoring for its data, (2) to assess the linguistic variance in text responses as a pivotal determinant of automatic scoring performance, and (3) to investigate the impacts of student and item characteristics to the linguistic variance.

New state-of-the-art benchmarks have been achieved for automatic scoring by using supervised deep learning and transformer architectures (Haller, Aldea, Seifert, & Strisciuglio, 2022; Whitmer et al., 2023). In contrast though, the current project focuses on operationally feasible approaches for international large-scale assessments and investigates linguistic variance in responses as a crucial factor for both the established and the new generation of automatic scoring systems (Haller et al., 2022), while still making use of pre-trained embeddings from transformer models.

Methods

Research Design

Instrument. In PIRLS, reading literacy is defined as the ability to understand and use written language required by society and/or valued by the individual (I. V. Mullis & Martin, 2019). The 2016 ePIRLS was an extension of PIRLS. It was offered in 2016 as an innovative assessment of online reading that was developed in response to the explosion of information available on the Internet. The 2016 ePIRLS was administered as computer-based assessment and simulated websites from the Internet, and students were required there to navigate the simulated websites to accomplish school-based research projects or tasks, since much online reading was done for the purpose of acquiring information. The 2016 ePIRLS assessment consisted of five tasks with each task lasting up to 40 minutes. Each student was asked to complete two of the tasks according to a specific rotation plan. The five 2016 ePIRLS tasks asked students to navigate through interconnected web pages containing both textual and visual information to complete school-like assessments about science and social studies topics. Each task involved approximately three different websites


timssandpirls.bc.edu/pirls2016/international-results/take-the-epirls-assessment/Mars/index-mars.html
TIME LEFT 38:32
PROGRESS
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
http://www.eightplanets.org/mars
The Solar System
8 PLANETS
Home Sun Mercury Venus Earth Mars Jupiter Saturn Uranus Neptune
Space Camp!
Take control of a trip to the stars!
Blast Off!
ePIRLS Class Project
SAVED
SAVE

INTRODUCTION TO MARS

The Red Planet

For hundreds of years, scientists have been interested in Mars. Because it is a planet next to Earth, perhaps it also has life?

At first, scientists could only study Mars through telescopes. They could see that Mars looked red. They found out that Mars is about half the size of Earth. Interestingly, until about 100 years ago, scientists thought that Mars had man-made canals. Because all living things on Earth need water, scientists thought there might be life on Mars!

As technology improved, we learned more about Mars. Now we know that Mars' color comes from reddish iron oxide on its surface. We learned that it is like a desert—very rocky and dusty. Also, Mars is farther away from the Sun than Earth, so it is much colder. In the winter, the temperature on Mars is about minus 100 degrees Celsius.

mercury

venus

earth

Mr. Webster
Now, click on Mars to learn more about this planet.

3.
100 years ago, why did scientists think there might be life on Mars?

Student

Figure 1. The 2016 ePIRLS Task *Mars* and Its Third Item.[1]

totaling about five to ten web pages. Reflecting the fact that online reading often involves sorting through more information than is necessary to achieve one's goal, the texts contained in the 2016 ePIRLS tasks average about 1,000 words in total (I. V. S. Mullis, Martin, Foy, & Hooper, 2017). Figure 1 shows the sample 2016 ePIRLS task *Mars*. The purpose of this item is classified as informational, and the cognitive process it measures is *making straightforward inferences*.

Out of 91 items in total in the assessment, more than half are in the CR format. There is a total of 51 CR items, two-thirds of which are dichotomous items and one-third of which are polytomous items. One of the CR items only involved the ordering of numbers and was thus not included in the present analysis, resulting in a total of 50 items. The CR items are spread over the five tasks and four comprehension processes, and the maximum score points vary across items, making it possible to examine the impacts of item characteristics.

Participants. The international PIRLS target population consists of students enrolled in the grade that represents four years of schooling (at least 9.5 years). The analyzed dataset

[1] https://timssandpirls.bc.edu/pirls2016/international-results/take-the-epirls-assessment/Mars/index-mars.html [2023-12-02]

involves 67,070 students from 15 countries/regions, who participated in the 2016 ePIRLS as part of PIRLS 2016 using 14 different test languages (s. Table 1).[2] The data covers a wide variety of levels of online informational reading proficiency, ranging from the United Arab Emirates (the lowest performing country with a mean scale score of 468) to Singapore (the highest performing country with a mean scale score of 588; I. V. S. Mullis et al., 2017).

Table 1
Number of Students per Test Language by Country/Region[a]

Country	**Language**	**Tag**	***n***	**Country**	**Language**	**Tag**	***n***
Chinese Taipei	T. Chinese	zh-TW	4,362	Singapore	English	en-SG	6,431
Denmark	Danish	da-DK	2,847	Slovenia	Slovenian	sl-SI	4,401
Georgia	Azerbaijani	az-GE	815	Sweden	Swedish	sv-SE	4,057
	Georgian	ka-GE	4,932	United Arab Emirates	Arabic	ar-AE	2,456
Ireland	English	en-IE	2,557		English	en-AE	1,774
Israel	Arabic	ar-IL	1,185	Dubai, UAE	Arabic	ar-AD	1,578
	Hebrew	he-IL	2,768		English	en-AD	5,995
Italy	Italian	it-IT	3,979		French	fr-AD[b]	154
Norway	Bokmål	nb-NO	3,495	Abu Dhabi, UAE	Arabic	ar-AAD	1,651
	Nynorsk	nn-NO[b]	331		English	en-AAD	2,458
Portugal	Portuguese	pt-PT	4,730	United States	English	en-US	4,114

[a] No data from Canada was included in the data available to the project.

[b] Excluded due to small sample size.

Treatment of Missing Data and Misclassification. As is common in the field of automatic scoring, we considered the manual score provided by human scorers as the gold standard. That is, human scores were the basis for both the training of automatic scoring models and its evaluation. This gold standard item-level score was included directly as a predictor, along with other student characteristics, in multilevel analyses and was also included indirectly in the derivation of outcome accuracy measures that address automatic scoring performance.

Despite the intense efforts made for ensuring the highest possible quality in large-scale

[2] See I. V. S. Mullis et al. (2017) for details on and constraints of the respective population coverage. For better legibility, we refer to participating parties as countries in this report rather than distinguishing countries and regions at each occurrence.

assessments, human scorers can introduce noise into scoring process through incorrect scoring if, for example, they are inattentive or the scoring guides do not cover a certain borderline response. These human scoring mistakes usually remain obscured if no double scoring is in place for a particular response and can sometimes be revealed by automatic scoring if very similar responses are scored differently.

The data contained different kinds of missing values and misclassifications. First, students can enter blank text responses. In the present study, these are considered as administered but intentionally omitted by students. Thus, the most obvious manual mistakes that can be dealt with is when blank responses receive a score other than specified in the scoring guides (i.e., *9* in PIRLS). Accordingly, we excluded blank responses that have been scored by humans as *0* ($n = 1693$), *1* ($n = 1934$), *2* ($n = 188$), or *3* ($n = 2$), totaling to 0.3 percent of all responses in the dataset. Second, responses for which no human score has been captured in the dataset were also excluded from analysis. Third, there were some non-empty, valid text responses misclassified as *9*. To keep the linguistic variance observed in those misclassified valid text responses, we recoded them to *0* (incorrect) responses. This treatment was irrelevant for computing *accuracy*, but affected quadratic weighted kappa–values because it involved weighting.

Automatic Scoring Models

International large-scale assessments present specific conditions for the automatic scoring of short text responses which influence the available methodological repertoire. They require the equivalent scoring across (i) a multitude of test languages, (ii) diverse item types and (iii) domains, and (iv) responses with diverse orthographic and grammatical irregularities, which can be attributed to the low stakes for the participants—or their age in the case of the 2016 ePIRLS. Furthermore, (v) the requirement for high scoring accuracy is critical, given the significant policy implications for participating countries.

In the present study, we adopted two approaches with different foci and thus strengths and weaknesses with respect to the aforementioned requirements: (1) fuzzy lexical matching, matching responses against a historic response database and (2) item-specific classifiers built by supervised learning of support vector machines with semantic sentence embeddings as the features. All classifiers, for both fuzzy lexical matching and the semantic ones, were built item-dependent for test language, and country (i.e., languages were not aggregated across different countries' data subsets).

Fuzzy Lexical Matching (FLM). Building upon PISA's Machine-Supported Coding System (MSCS; Yamamoto, He, Shin, & von Davier, 2017), we implemented a fuzzy lexical matching (FLM) system of text responses in the 2016 ePIRLS to historic responses. These historic responses stem from the same items and had been scored previously by humans. In contrast to the MSCS, we did not require responses to match exactly character by character and case-sensitively but instead normalized texts by means of established preprocessing techniques from natural language processing. This makes the matching somewhat fuzzier. That is, a new text response such as *"The dog runs quickly..."* would not be matched with a historic response such as *"DOg run quick"* under the exact-matching MSCS but would be considered identical with our fuzzy matching procedure. Preprocessing both responses to *"dog run fast"* allows them to be matched at a more simplified (and less linguistically diverse) string level. If matched, the score of the historic response would be propagated to the new response. The following normalization steps were applied for this: First, (1) redundant whitespaces in the texts, such as double spaces between words, were removed. Next, (2) punctuation was removed, and (3) responses were converted to their lower case, followed by (4) stemming, (5) ignoring word order (bag of words) and (6) removing diacritics. Some of the steps are interdependent, e.g. words must be converted to their lower case to match the stop word list. Other steps are language- and resource-dependent, such as stop word removal and stemming rules. Moreover, score propagation was slightly adapted compared to the MSCS. The MSCS requires at least five matching responses with identical scores for propagation. This means that for a new response to be scored automatically, it must match exactly at least five responses in the historical database. To propagate scores within a matched group using FLM, on the other hand, required a minimum of three matching responses with at least 92 percent consistent scores[3] and an absolute maximum of five deviating scores in that response group. This takes minor human misclassifications into account while still adhering to high requirements regarding accuracy. For evaluation

[3] An inter-rater agreement of 92 percent is the expected domain-level standard across all items in a domain in PISA (OECD, 2023).

purposes, we simulated the historic database to consist of the dataset described above.

Supervised Classifiers Based on Semantics (XLM-R & SVM). In contrast to the character-based matching described above, we additionally implemented a semantics-based approach. To achieve this, we trained item-, country-, and language-specific classifiers using supervised learning and response vectors from pre-trained embeddings, representing the responses' semantics, as its features. This was achieved using XLM-RoBERTa (XLM-R) in its base version (Conneau et al., 2020), which is a pre-trained deep neural network with a transformer architecture and cross-lingual representations of about one hundred languages. XLM-R offers the potential to be used for cross-lingual transfer tasks, and also performs well on low-resource languages (Conneau et al., 2020). When using a text response as the input, the model's last network layer can be considered as its semantic representation. Specifically, this way, responses are represented by 768-dimensional vectors and responses with vectors pointing in similar directions (i.e., with similar values) are considered semantically close. Importantly, the model's attention mechanism provides context, allowing words to be disambiguated by the context of the response. Thus, the text response is passed to the model as a string, to be tokenized and vectorized based on its word order. Given the extracted semantic response vectors as features, we trained support vector machines with a radial kernel to build item-, country-, and language-specific classifiers.

Advantages and Disadvantages of Both Approaches. The advantage of the fuzzy lexical matching is that it works straightforward across all test languages, reproducible, and transparent. Its disadvantage is that it is not applicable to any unmatched response, which becomes very prevalent in items with medium and high linguistic variance in the text responses. We report this as the degree of efficiency in the results. Also, while the basic concept of string-matching makes the algorithm language-agnostic, the coverage of automatically scorable responses (i.e., added value in efficiency) does vary by language due to linguistic phenomena and characteristics of the country-specific dataset (e.g., frequency of incorrect, correct, and blank responses), making the equivalence of its cross-lingual applicability an empirical matter. On the other hand, the supervised approach is known to provide scoring accuracy close to that of human performance. One advantage of these classifiers is that they can score every new response, unlike fuzzy lexical matching, and XLM-R provides sufficiently stable representations for most of the fourteen test languages in the analyzed data, which results in different human-machine agreements. Additionally other inherent language-dependent characteristics and variance also play an important role, affecting the reliability of response representations and success of the machine learning model in classifying unseen instances. However, scoring accuracy can vary greatly across items and is an empirical question for given test languages, items, and training data size for a particular test language.

Evaluation Metrics. We report accuracy and quadratic weighted kappa to evaluate scoring performance (for FLM and XLM-R & SVM) and additionally efficiency for FLM. *Efficiency* measures the extent to which human scoring effort is reduced by using fuzzy lexical matching and score propagation. This measure is not applicable to supervised classifiers based on semantics, as they score all text responses without exception. *Accuracy* quantifies the percentage of exact agreement between machine-predicted scores and human-assigned scores. We evaluate the accuracy measure for both fuzzy lexical matching and supervised classifiers based on semantics. *Quadratic weighted kappa (QWK)* (Cohen, 1960) similarly captures this agreement but corrects for agreement by chance. We use QWK as a supplementary measure for supervised classifiers based on semantics. For evaluating the accuracy and quadratic weighted kappa of automatic scoring performance, a 5-fold cross-validation was conducted, repeated five times. A split into five folds allows a sufficient representation of skewed scores within single folds.

Assessing Linguistic Variance

In order to assess linguistic variance of responses, we considered basic measures that operate on a set of responses to one item as a corpus. We followed Horbach and Zesch (2019) and used a variant of the **type-token ratio (TTR)**, *STTR* (sampled type-token ratio). TTR is normally used to measure the lexical variance within a single text by dividing the number of different tokens—that is, types—in a text by the overall number of tokens. In our analyses, however, we aimed to measure the variance operationalized through lexical diversity across a set of responses and languages. Therefore, we first built a pseudo-document by concatenating tokens drawn randomly from the entire pool of responses for a certain subset.

As TTR values are known to be affected by text length (Koizumi, 2012), we constructed the pseudo-documents by sampling a fixed number of tokens (100) from the entirety of responses of a particular subset of interest and repeated this sampling several times (5000) in order to avoid artifacts from random sampling (cf. Horbach and Zesch (2019)). This way, STTR values were comparable across items and languages. The relatively small number of sampled tokens and the high number of iterations allowed STTR to be used even for items with very low linguistic diversity and small subsamples, as required by the multilevel analysis described in the following section.

Multilevel Analyses

Multilevel analyses can be useful in decomposing the sources of variance by different clusters. This allows for a better understanding of the influences of student- and item-related factors on linguistic variance simultaneously. The models described below aim to analyze the sources of linguistic variance in text responses with respect to item- and student-related characteristics. Additionally, they aim to investigate the resulting relationship on the performance of automatic scoring.

STTR is calculated at the level of pseudo-documents consisting of several responses written by students who responded to that item from a certain language-country group. The data follows a multilevel structure, depending on how pseudo-documents are constructed. There are two ways to construct a pseudo-document for analysis. The first involves collecting text responses per item in each language-country group, which allows for investigation of item-related characteristics. The second involves breaking down text responses per student group, item, and language-country group, allowing for simultaneous examination of the effects of item- and student-related characteristics.

We analyzed two sets of STTR values. The first set ($STTR^{(1)}_{ig}$) was calculated for each item per language-country group, while the second set ($STTR^{(2)}_{pig}$) was calculated at the level of the combination for student-characteristics, including gender and L1/L2 speaker status. Regarding the first set of STTR values ($STTR^{(1)}_{ig}$), the pseudo-document is constructed for each item (level 1; denoted as i) per language-country group (level 2; denoted as g). Membership to a certain language-country group can be reflected as a random effect, allowing for variation across different language-country groups. As an extension, item characteristics, such as the maximum score points for an item and its position within the passage, can be specified as fixed effects (details below). It is important to note that $STTR^{(1)}_{ig}$ values for two-level models are directly extracted from the student delivery system and were not yet matched to the public use file, making them more abundant. In total, $STTR^{(1)}_{ig}$ values consisted of 1,000 records for 50 items grouped into 20 language-country clusters.

On the other hand, if the pseudo-document is constructed at a more granular level, taking into account student characteristics as in the $STTR^{(2)}_{pig}$, a lower cluster was added, resulting in a three-level structure: one for students grouped by gender, L1/L2 speaker status, and their scores on that item (level 1; denoted as p), one for items (level 2; denoted as i), and one for language-country groups (level 3; denoted as g). Similarly, characteristics of students for that combination of variables can be specified as fixed effects, while items and language-country groups can be considered as random effects. Not only student characteristics (level-1 covariates), but also item-related characteristics can be included as level-2 covariates. To accomplish this, we calculated the $STTR^{(2)}_{pig}$ values by taking into account the information about students' characteristics that was recorded in the public use file, matching student identifiers. Pseudo-documents were created for each combination of student characteristics (gender, L1/L2 speaker status, and corresponding scores for each item), items, and language-country groups. Cases where the number of text responses used for constructing pseudo-documents was less than 30 were excluded. Therefore, the final dataset used for multilevel analyses comprises 4,863 cases, and the range of text responses used for constructing pseudo-documents was between 31 and 426. In total, we fit the following four models. For fitting multilevel models, the package `lme4` (v. 1.1-35.1; Bates, Mächler, Bolker, & Walker, 2014) in `R` 4.2.1 was used. In the results, restricted maximum likelihood (REML) estimates are reported.

Model M1: Variance-components models. The most basic approach is to model the relationship of item characteristics and STTR without any covariates. This involves decomposing the total variance by the respective clusters: all realizations of language-country group g for $STTR^{(1)}_{ig}$ as well as language-country group g and item i for $STTR^{(2)}_{pig}$, respectively. Equation (1) denotes the unconditional random-intercept model where item-level STTR values are nested in language-country groups. In this two-level model, the total variance in STTR is split into two error components: ζ_g, which is shared between items of the same language-country groups, and ϵ_{ig}, which is unique for each item.

Next, relatively finer STTR values for groups of students are nested in items and in language-country groups, as formulated in Equation (2). This three-level model suggests that STTR values for the same language-country group can be correlated, noted as the shared level-3 random intercept $\zeta_g^{(3)}$. Conditional on $\zeta_g^{(3)}$, STTR values for the same item are not independent but correlated, and they depend on the shared level-2 random intercepts, $\zeta_{ig}^{(2)}$. In this model, the level-1 variance θ can be interpreted as the between-students' characteristics, within items, and within-language-country variance. The level-2 variance $\psi^{(2)}$ is the between-items, within-language-country variance. And the level-3 variance $\psi^{(3)}$ is the between language-country variance. All three error components are uncorrelated across language-country groups, the level-2 random intercepts and level-1 residuals are uncorrelated across items, and the level-1 residuals are uncorrelated across student characteristics.

$$\begin{aligned} STTR_{ig}^{(1)} &= \beta + \zeta_g + \epsilon_{ig}, \\ \zeta_g &\sim \mathcal{N}(0, \psi), \\ \epsilon_{ig} &\sim \mathcal{N}(0, \theta). \end{aligned} \tag{1}$$

$$\begin{aligned} STTR_{pig}^{(2)} &= \beta + \zeta_{ig}^{(2)} + \zeta_g^{(3)} + \epsilon_{pig}, \\ \zeta_{ig}^{(2)} &\sim \mathcal{N}(0, \psi^{(2)}), \\ \zeta_g^{(3)} &\sim \mathcal{N}(0, \psi^{(3)}), \\ \epsilon_{pig} &\sim \mathcal{N}(0, \theta). \end{aligned} \tag{2}$$

Since our focus is on decomposing linguistic variance by those clusters, intra-class correlations (ICCs) are reported as indicators of how much of the total variability is explained by group membership. More specifically, ICC by language-country groups (γ_g) resulting from the two-level model is written in Equation (3). For three-level models, two types of ICCs can be considered as written in Equation (4). For the same language-country group g, different items i and i' and different student characteristics p and p', the ICC becomes

$\rho(g)$. For the same language-country group g and for a given item i, ICC becomes $\rho(i, g)$.

$$\rho = \frac{\psi}{\psi + \theta}. \tag{3}$$

$$\begin{aligned} \rho(g) &= \frac{\psi^{(3)}}{\psi^{(2)} + \psi^{(3)} + \theta}, \\ \rho(i, g) &= \frac{\psi^{(2)} + \psi^{(3)}}{\psi^{(2)} + \psi^{(3)} + \theta}. \end{aligned} \tag{4}$$

Model M2: Random-intercept STTR models with item characteristics. Next, expanding the variance-components models in M1, we include the following item characteristics in M2. The following item characteristics are considered.

- **Process**: four cognitive processes of comprehension (I. V. S. Mullis & Martin, 2015)
  - Focus On and Retrieve Explicitly Stated Information [F]
  - Make Straightforward Inferences [M]
  - Interpret and Integrate Ideas and Information [I]
  - Evaluate and Critique Content and Textual Elements *(reference group)*
- **Passage**: school-based online reading tasks, each of which involves 2–3 different websites, totaling to 5 to 10 web pages, together with a series of comprehension questions based on the task (Martin, Mullis, & Foy, 2016)
  - Mars [M]
  - Rainforests [R]
  - The Legend of Troy [T]
  - Zebra and Wildebeest Migration [Z]
  - Dr. Elizabeth Blackwell *(reference group)*
- **Position**: position of the item within the passage[4]
- **Maximum Points**: maximum score for an item's response, ranging from 1 to 3
- **Difficulty**[5]: proportion of incorrect responses per item as a proxy of item difficulty, with higher values indicating more difficult items

[4] We assume that the impact of the within-passage position is still important, as students took two passages according to the rotated matrix sampling design (Martin et al., 2016), which cancels out the impacts of between-passage position.

[5] Retrieved from https://timssandpirls.bc.edu/pirls2016/international-database/index.html [2023-12-01]

The inclusion of item characteristics results in Equation (5), where x_{kig} is the k^{th} item-related predictor and β_k indicates the k^{th} regression coefficient for the associated predictor. In this model, item-characteristics are level-1 predictors.

$$\begin{aligned} STTR_{ig}^{(1)} &= \beta_1 + \zeta_g + \beta_2 x_{2ig} + ... + \beta_k x_{kig} + \epsilon_{ig}, \\ \zeta_g &\sim \mathcal{N}(0, \psi), \\ \epsilon_{ig} &\sim \mathcal{N}(0, \theta). \end{aligned} \tag{5}$$

Similarly, for the three-level model, the item characteristics can be specified as well. Unlike the two-level model, item-characteristics are included as level-2 predictors, which can be written as Equation (6).

$$\begin{aligned} STTR_{pig}^{(2)} &= \beta_1 + \zeta_{ig}^{(2)} + \zeta_g^{(3)} + \beta_2 x_{2ig} + ... + \beta_k x_{kig} + \epsilon_{pig}, \\ \zeta_{ig}^{(2)} &\sim \mathcal{N}(0, \psi^{(2)}), \\ \zeta_g^{(3)} &\sim \mathcal{N}(0, \psi^{(3)}), \\ \epsilon_{pig} &\sim \mathcal{N}(0, \theta). \end{aligned} \tag{6}$$

The comparison of M1 and M2 for each two-level model and three-level model reveals how influential the item characteristics are for explaining the linguistic variance.

Model M3: Random-intercept STTR models with student characteristics. Another way to expand the variance-components models in M1 is to specify student characteristics to investigate their impact. It is important to note that student characteristics cannot be included in the two-level model, since the lowest unit of STTR was at the item-level. Therefore, M3 is only considered for the three-level model with the following student characteristics.

- **Gender**: 1 for female *(reference group)* and 2 for male
- **L1/L2**: L1 refers to the student's first language, while L2 refers to their second language or the language they are currently learning. This variable was constructed in the following way by utilizing the information collected and provided in the public use file database. First, we started with the questionnaire item `ASBG03: How often do you speak <language of test> at home?` by recoding the first two options (1

and 2) as 0, representing L1 *(reference group)*, and the latter two options (3 and 4) as 1, representing L2. Further, if the response to `ASBG03` was not provided, `ASBH17: How often does your child speak <language of test> at home?` was additionally used with the same recoding scheme. Below are the response options for `ASBG03` and `ASBH17`, respectively.

1. `I always speak <language of test> at home` | `Always`
2. `I almost always speak <language of test> at home` | `Almost always`
3. `I sometimes speak <language of test> and sometimes speak another language at home` | `Sometimes`
4. `I never speak <language of test> at home` | `Never`

- **Score**: 0 for incorrect responses for all items, and 1 for correct response for dichotomous items or partially correct responses for polytomous items. Maximum score points were up to 2 or 3 for polytomous items.

The model formulation for M3 7 is the same as M2 in Equation 6 except that explanatory variables are now level-1 student characteristics. For the ease of notation, level-1 predictors (student characteristics) are denoted as z_{pig} with the associated regression coefficients of γs. The comparison of M1 and M3 for three-level model reveals how influential the student characteristics are for explaining the observed linguistic variance in text responses.

$$\begin{gathered} STTR^{(2)}_{pig} = \beta_1 + \zeta^{(2)}_{ig} + \zeta^{(3)}_{g} + \gamma_2 z_{2pig} + ... + \gamma_k z_{kpig} + \epsilon_{pig}, \\ \zeta^{(2)}_{ig} \sim \mathcal{N}(0, \psi^{(2)}), \\ \zeta^{(3)}_{g} \sim \mathcal{N}(0, \psi^{(3)}), \\ \epsilon_{pig} \sim \mathcal{N}(0, \theta). \end{gathered} \tag{7}$$

Model M4: Random-intercept STTR models with item- and student-related characteristics. The most comprehensive model specified in this study utilizes both item-related (x_{ig}) and student-related (z_{pig}) characteristics. M4 can be written as Equation (8).

$$\begin{gathered}
STTR_{pig}^{(2)} = \beta_1 + \zeta_{ig}^{(2)} + \zeta_{g}^{(3)} + \beta_2 x_{2ig} + ... + \beta_k x_{kig} + \gamma_2 z_{2pig} + ... + \gamma_k z_{kpig} + \epsilon_{pig}, \\
\zeta_{ig}^{(2)} \sim \mathcal{N}(0, \psi^{(2)}), \\
\zeta_{g}^{(3)} \sim \mathcal{N}(0, \psi^{(3)}), \\
\epsilon_{pig} \sim \mathcal{N}(0, \theta).
\end{gathered} \tag{8}$$

Comparisons of M2 vs. M4 or M3 vs. M4 for the three-level models reveal how much the other set of characteristics contribute to explain the total variance of STTR.

Results

This section first reports performance of automatic scoring for the employed Fuzzy Lexical Matching (FLM) and scoring with XLM-R in terms of reduction in manual effort (i.e., efficiency) and agreement with human scores (i.e., accuracy), respectively. Next, it presents the relationship between linguistic variance in text responses and item and student characteristics. Finally, it reports the relationship of automatic scoring performance—again, in terms of efficiency and accuracy—with linguistic variance, item, and student characteristics.

Performance of Automatic Scoring

In terms of agreement between human and machine scores, the performance of automatic scoring varied mostly across items and partly across country-language groups. The highest agreement was found for Item B03 ($\kappa = .966$) and the lowest agreement was found for Item M20 ($\kappa = .476$), which is considered a range of *fair to good* up to *excellent* agreement beyond chance (Fleiss, 1981). Out of a total of 50 items, 36 show an agreement of at least $\kappa \geq .650$, 22 items of at least $\kappa \geq .750$, and 11 items of at least $\kappa \geq .850$. The average agreement across all items and country-language groups is $\kappa = .745$, which is at the upper boundary of *fair to good agreement* beyond chance.

With respect to country-language groups (Table 2), the highest agreement between human and machine was observed for zh-TW ($\kappa = .814$), followed by en-SG ($\kappa = .804$), and en-AD ($\kappa = .795$). The lowest agreement on average was observed for az-GE ($\kappa = .706$).

When comparing the automatic scoring by XLM-R for different countries that use the same test language, agreements were similarly high. Both Arabic and English were used as the test language in multiple countries. For Arabic, agreements ranged between $\kappa = .720$ (ar-IL) and $\kappa = .772$ (ar-AD), and for English, they ranged from $\kappa = .723$ (en-IL) to $\kappa = .804$ (en-SG).

It is important to note that the overall agreement between automatic and human scoring is higher when including blank responses, as their scoring is straightforward. However, they are typically included in reporting inter-rater agreement for indicating the effective operational impact automatic scoring would have. Blank responses constitute a non-negligible portion of the total responses, varying substantially across languages and countries (10.6% on average, ranging from 2.3% for en-SG to 33.6% for az-GE). With blank responses not including any linguistic variance, the efficiency of the FLM approach is affected similarly.

FLM reaches its highest efficiency (again, see Table 2) for az-GE (37.9%) and zh-TW (32.0%), while it is lowest for he-IL (17.1%) and for en-IE (17.7%). Essentially, the manual scoring effort can be reduced by about 17% for these when applying FLM. The reduced efficiency, compared to the other language-country groups, is likely attributable to significant realization variance (lexical diversity, misspelling, etc.) or conceptual heterogeneity in responses as their percentage of blank responses is relatively low.

When responses exhibit a wide variation, the threshold for propagating scores automatically is not passed, leading to the need for manual scoring. Moreover, low levels of agreement in FLM can either indicate a negative impact of the text preprocessing, human scoring inconsistencies, or both. That is, on the one hand, preprocessing can sometimes remove relevant information for determining a response's score. On the other hand, closer examination of the responses revealed a significant degree of inconsistencies in the human scores, where different scores were assigned to identical responses.

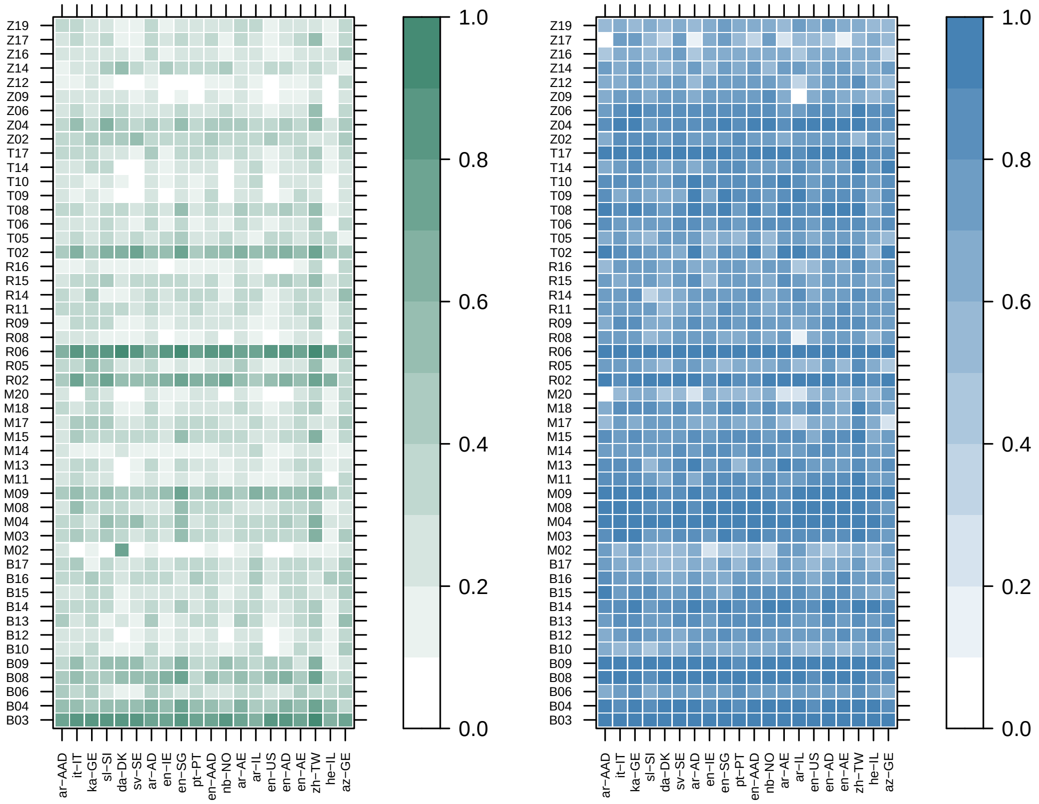
As an example, the prevalence of misclassifications was particularly pronounced in Item B10 of Passage [B] for en-AD. The item received a total of $n = 2,375$ responses, out of which 146 were blank. Four of these blank responses were not assigned the intended code 9 (i.e., invalid). Three blank responses were erroneously scored as *0 – No Comprehension* and two as *1 – Partial Comprehension*. More significantly, a non-negligible portion of identical text responses had inconsistent scores. For instance, out of 48 responses, reading `she kept`

Table 2
Automatic Scoring Performance

	Fuzzy Lexical Matching			**XLM-R**	
	Efficiency	Accuracy	κ	Accuracy	κ
	%	%		%	
ar-AAD	26.6	98.6	.935	92.7	.747
ar-AD	28.5	98.2	.914	91.1	.772
ar-AE	27.7	98.0	.964	91.0	.784
ar-IL	25.8	98.0	.853	87.8	.720
az-GE	37.9	98.9	.969	90.4	.706
da-DK	19.1	93.0	.722	84.6	.710
en-AAD	26.7	97.6	.947	88.4	.790
en-AD	23.3	97.6	.930	87.1	.795
en-AE	27.5	98.6	.914	88.7	.772
en-IE	17.7	96.3	.812	85.1	.723
en-SG	27.2	97.2	.926	88.3	.804
en-US	17.8	96.9	.903	84.4	.742
he-IL	17.1	97.9	.869	85.1	.718
it-IT	26.4	97.9	.966	86.8	.762
ka-GE	27.5	97.0	.937	88.9	.787
nb-NO	17.8	94.6	.761	86.2	.742
pt-PT	22.1	98.2	.959	86.9	.757
sl-SI	24.9	96.1	.899	84.7	.718
sv-SE	18.8	94.6	.808	86.6	.743
zh-TW	32.0	96.0	.919	89.7	.814

`working hard as a general doctor`, 33 were scored in accordance with the scoring guide (*1 – Partial Comprehension*) and 15 as *2 – Full Comprehension.* Consequently, and with this only being one type of inconsistency, both training and and evaluation of automatic scoring were impeded.

Figure 2 presents an overview of both FLM's efficiency and XLM-R's scoring performance in terms of quadratic weighted kappa. Obviously, both measures of automatic scoring performance varied widely from item to item, although though the overall performance was acceptable on average. For the operational use of automatic scoring, certain items could be identified for which it might be sufficiently mature, whereas others would still require manual scoring to a smaller or even larger extent. This highlights the importance of understanding the factors that determine the performance of automatic scoring. Consequently, the following analyses first assess the linguistic variance prevalent in text responses and



(a) Efficiency of FLM

(b) QWK of XLM-R

Figure 2. Performance of Automatic Scoring

then examine its relationships with item and student characteristics as well as, ultimately, with automatic scoring performance.

Descriptive Statistics of Linguistic Variance

Instrument-Related Factors ($STTR_{ig}^{(1)}$). Linguistic variance measured through STTR is illustrated in four figures, namely by language-country group, by item, by process, and by an item's position within its passage. Firstly, as shown in Figure 3, students in ar-AAD showed the largest median STTR value (.711) while those in zh-TW showed the smallest median STTR value (.445) across items. There were two languages, Arabic and English, that were used in different regions or countries. Within the Arabic language, AAD stood out as having the largest linguistic variance, followed by AD and AE with very similar

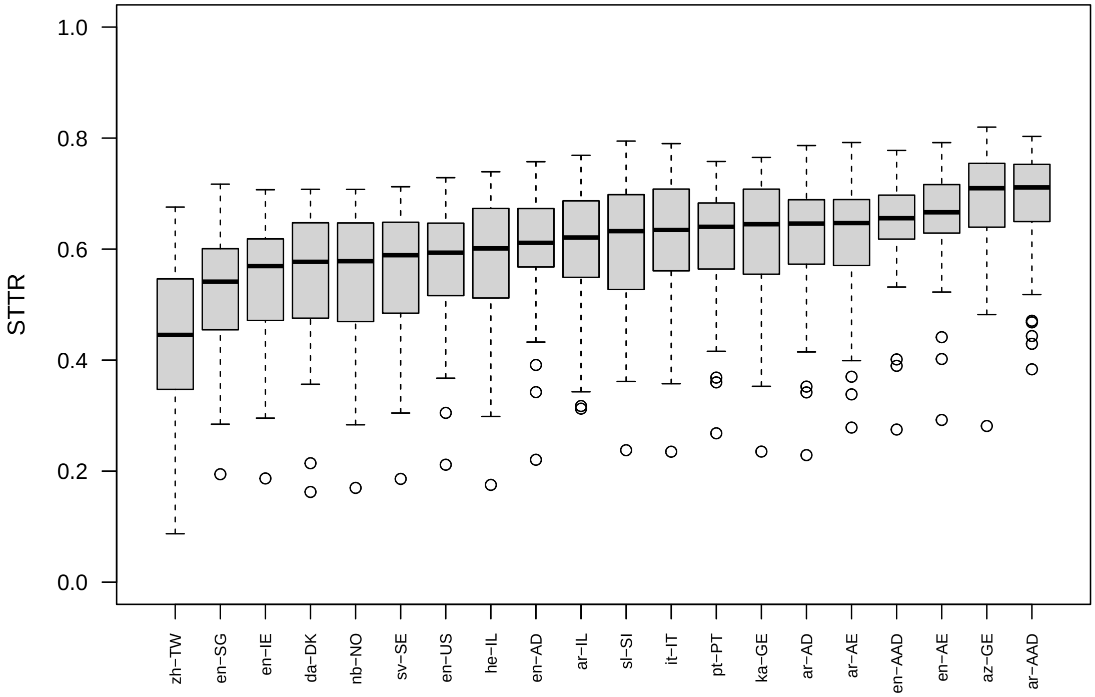


Figure 3. Linguistic Variance by Language-Country Group ($STTR_{ig}^{(1)}$)

median STTR values (.646), and followed by IL with a somewhat smaller median STTR value (.621). For English, there were six regions/countries in total. In contrast to Arabic, AE and AAD showed the largest median STTR value (.666 and .656) followed by AD (.611), US (.594), IE (.569), and SG (.541). Ranking-wise, means of STTR showed the same pattern. Standard deviations of STTR values iterated over 5000 times for each item per language-country group were very consistent with .10 to .11.

Second, Figure 4 shows the distribution of STTR per item across language-country groups. Comparing the two figures (3 and 4) shows that most of the systematic variation is at the item level, as the range of within-item STTR is much smaller compared to language-country specificity. This is in line with conceptual assumptions, since the same set of translated items elicits similar vocabulary spaces across countries and language groups. Interestingly, STTR and item difficulty were moderately correlated (r = .529). That is,

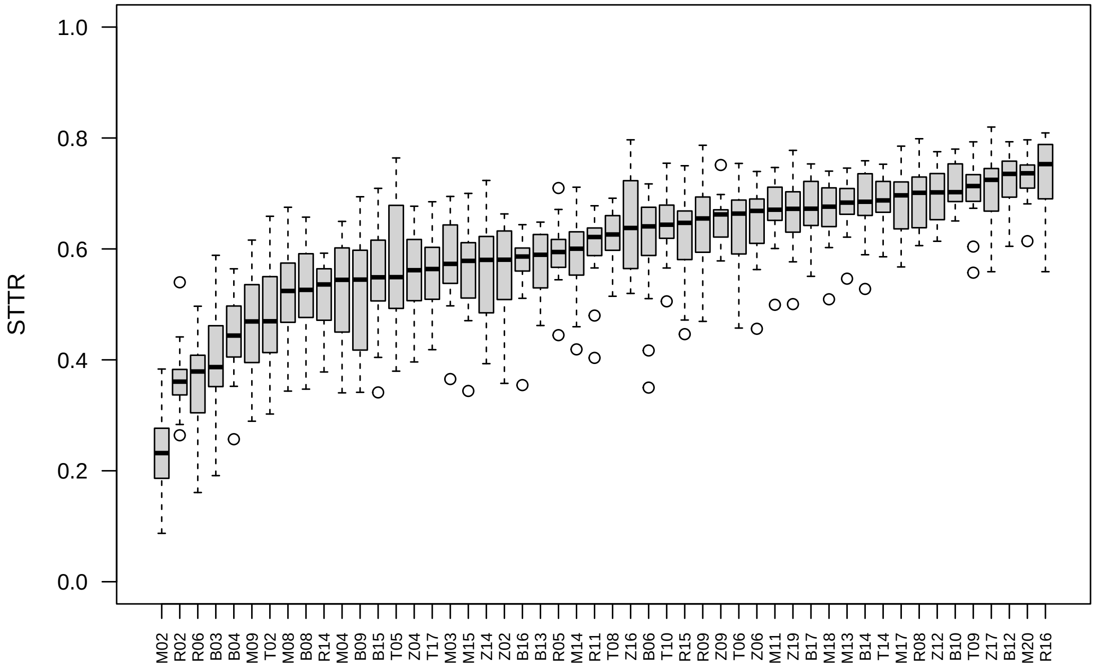


Figure 4. Linguistic Variance by Item ($STTR_{ig}^{(1)}$)

more difficult items tended to elicit relatively more heterogeneous responses from students, as indicated by greater linguistic variance in general.

Figure 5 illustrates that the median STTR value was lowest for “Focus On and Retrieve Explicitly Stated Information” [F] at .472. The other three processes had rather similar median STTR values: .621 for “Make Straightforward Inferences” [M], .673 for “Interpret and Integrate Ideas and Information” [I], and .662 for “Evaluate and Critique Content and Textual Elements” [E].

Furthermore, as Figure 6 shows, median STTR values tended to become larger as students progressed within the passage. Per passage, the correlation between the sequence of the item and the median STTR values ranged from r = .267 in *The Legend of Troy*, r = .363 in *Zebra and Wildebeest Migration*, r = .484 in *Dr. Elizabeth Blackwell*, r = .544 in *Rainforests*, and r = .659 in *Mars*. This could be a design feature of the 2016 ePIRLS

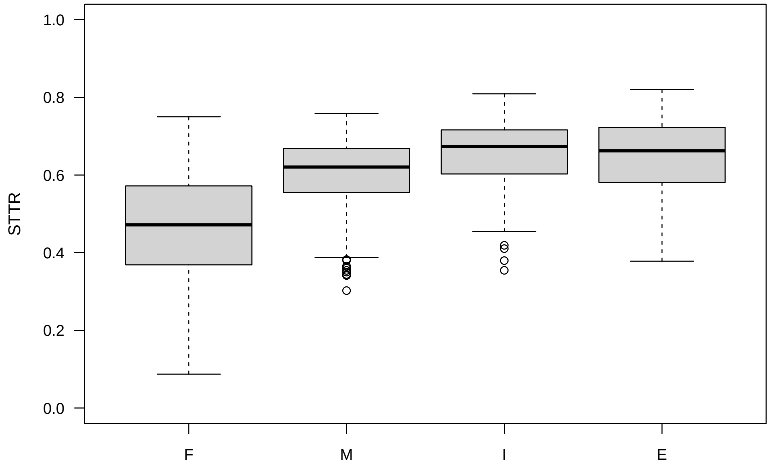


Figure 5. Linguistic Variance by the Cognitive Process Elicited by an Item ($STTR_{ig}^{(1)}$)

assessment, if item developers deliberately placed more complex thinking items in the later position within the passage.

Student-related Factors ($STTR_{pig}^{(2)}$). Distribution of linguistic variance was also examined by student characteristics: gender, L1/L2 speaker status, and the score (see Figure 7 and Figure 8). For student characteristics, $STTR_{pig}^{(2)}$ was used. By gender, median values of linguistic variance were almost identical: .589 for girls and .608 for boys, with a difference of .019. Concerning L1/L2 speaker status, median values of linguistic variance were almost identical as well between the native speaker (L1; .604) compared to the L2 speaker (.585) with a difference of .019.

Finally, when considering the score of responses and an item's maximum points (refer to Figure 8), it is evident that linguistic variance had the highest median STTR value in incorrect responses compared to correct responses. In fact, correct responses generally exhibited smaller STTR values for the given maximum points of items.

Decomposing Linguistic Variance in Multilevel Analyses

The results of multilevel analyses for two sets of STTR values are presented in this section. Specifically, $STTR_{ig}^{(1)}$ is used for two-level models, while $STTR_{pig}^{(2)}$ is used for three-

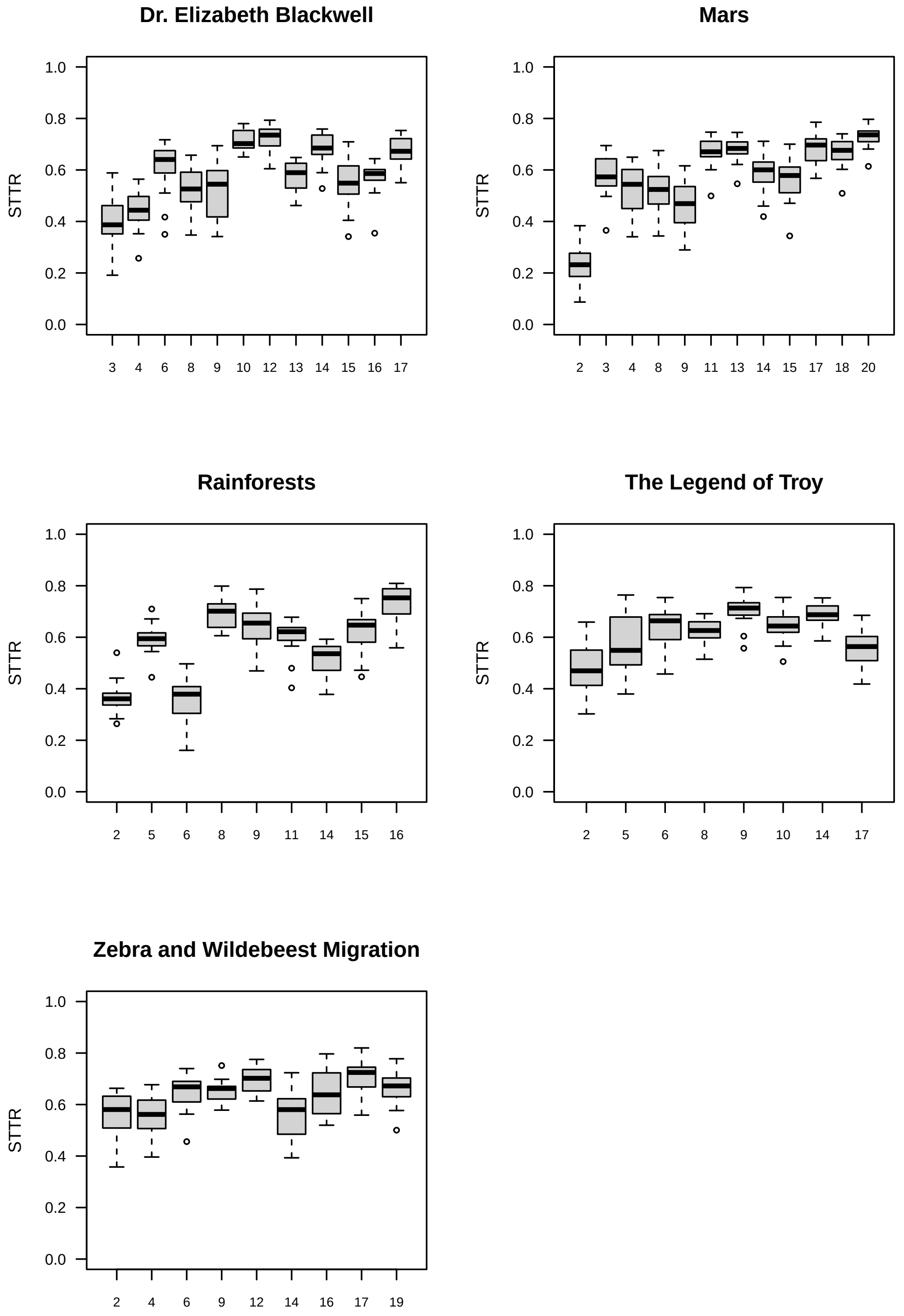


Figure 6. Linguistic Variance by Item Position in the Passage ($STTR_{ig}^{(1)}$)

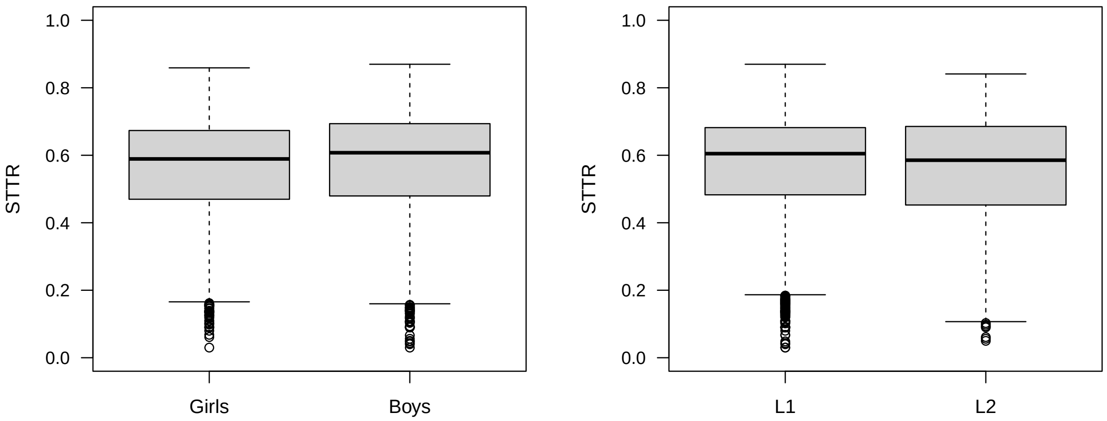


Figure 7. Linguistic Variance by Gender (left) and L1/L2 status (right; $STTR^{(2)}_{pig}$)

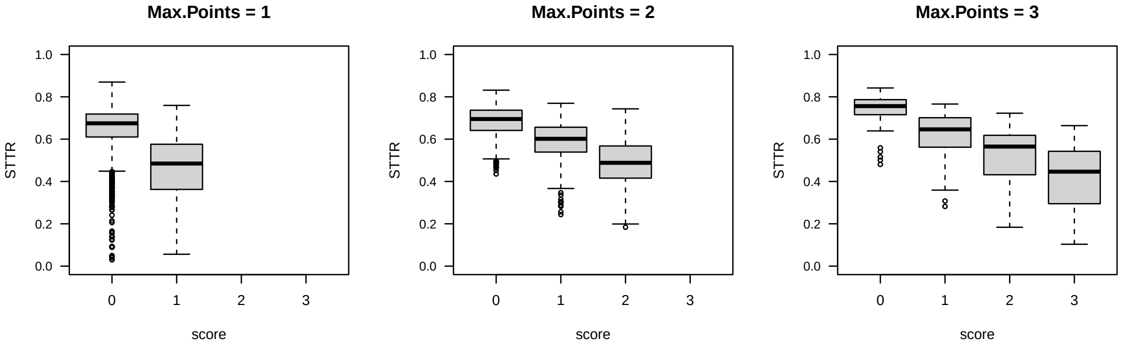


Figure 8. Linguistic Variance by Score Level and an Item's Maximum Points ($STTR^{(2)}_{pig}$)

level models. Two-level models are employed to examine the relationship between linguistic variance and item characteristics, taking into account the possibility of average linguistic variance varying across language-country groups. Three-level models are extended versions used to examine the relationship between linguistic variance and student characteristics. These models take into account that the average linguistic variance may vary across items and language-country groups.

Results of the Two-level Models. As a baseline model, M1 is useful for decomposing the total variance into different clusters. The ICC of M1 for the two-level model was $\rho = .199$, indicating that 19.9% of the total linguistic variance is explained by language-country groups.

Table 3
Effects of Item Characteristics on Linguistic Variance

	Model 1		**Model 2**	
Predictors	Estimate	SE	Estimate	SE
(Intercept)	.594***	.013	.416***	.019
Process [F]			–.078***	.010
Process [M]			.010	.008
Process [I]			.041***	.007
Passage [M]			.003	.006
Passage [R]			.016*	.008
Passage [T]			.020**	.008
Passage [Z]			.049***	.007
Position			.004***	.001
Max. Points [2]			.025***	.006
Max. Points [3]			–.003	.011
Difficulty			.295***	.017
Random Effects				
ψ	.003		.003	
θ	.013		.005	
Marginal R^2	.000		.501	
Conditional R^2	.199		.709	

*** for $p < .001$, ** for $p < .01$, and * for $p < .05$

The estimated regression coefficients using $STTR^{(1)}_{ig}$ as the dependent variable are shown in Table 3. For M2, item-related characteristics were specified as main effects only at this point. At the significance level of $\alpha = .001$, the variables Process [F], Process [I], Passage [Z], Position, Maximum Points [2], and Difficulty were statistically significant, controlling for the other covariates. Passages [R] and [S] were found to be significant at a more liberal significance level. More specifically, regarding *process*, the expected STTR is .078 lower for "Focus on and retrieve explicitly stated information" items compared to "Evaluate and critique content and textual elements" items, controlling for other covariates. Given that [E] items are supposed to require more complex thinking to articulate and respond than [F] items, the observation of less linguistic variance for [F] items is consistent with expectation. Another process, "Interpreting and Integrating Ideas and Information" items, which are also used for more complex responses, tend to show more linguistic variance compared to [E] items, as much as .041 on average.

In terms of *process*, the items “Rainforests”, “The Legend of Troy”, and “Zebra and Wildebeest Migration” showed larger linguistic variance on average than the items “Dr. Elizabeth Blackwell”, while the items “Mars” showed similar levels of linguistic variance, controlling for other covariates. Next, as shown in the 6, the STTR was estimated to increase on average by as much as .004 when the item is located in the later *position*. Regarding *maximum score points*, the linguistic variance for 3-score point items was not statistically different from the 1-score point items, while 2-score point items tended to show a slightly larger linguistic variance on average. This could be due to the fact that most of the items are 1-score point items (36 out of 50) and there are only three 3-score point items. Finally, the estimated regression coefficient for *Difficulty* shows that more difficult items are associated with higher STTR by an average of .295, controlling for other covariates. This result is not surprising given that items that require students to provide longer responses may be more difficult items.

When these item characteristics were specified as main effects in M2, the marginal $R^2 = .501$. This suggests that the linguistic variance explained by the item characteristics is 50.1%. Together with the variance explained by the language-country groups, the total variance explained by the fixed and random effects is 70.9%, as shown in the conditional R^2.

Results of the Three-Level Models. Similar to the two-level models, M1 was used to decompose the total variance into different clusters. In particular, two types of ICCs obtained from M1 of the three-level model showed that $\rho(g) = .093$ and $\rho(i, g) = .482$. These values indicate that 9.3% of the total variation in linguistic variance is explained by between-language/country groups, and 48.2% is explained by between-item within-language/country groups.

Table 4 reports estimated regression coefficients for three-level models with $STTR^{(2)}_{pig}$ as the dependent variable. Consistent with the two-level model, M2 included only item-related characteristics as main effects. In addition, for the three-level model only, M3 specified only student-related characteristics as main effects, while M4 specified both student-related and item-related effects as the most general model.

Regarding M2, the pattern of item-related effects is consistent with the two-level model.

The only difference is that passages [R] and [T] are no longer significantly different from the reference passage [D], and the effects of maximum score points were reversed when STTR values were calculated taking into account student characteristics. However, the effects of Process, Position, and Difficulty are still in line with the expectation as well as the empirical results of the two-level model. Specifying these level-2 item characteristics explained up to 29.0% of the linguistic variance, as shown in the marginal R^2.

In M3, only student characteristics were specified as predictors. In line with conceptual assumptions, significantly less linguistic variance was observed for *correct responses* compared to incorrect responses. On average, a score of 1 tended to show .140 less STTR values, a score of 2 tended to show .199 less STTR values, and a score of 3 tended to show .316 less STTR values compared to incorrect responses (score of 0), controlling for gender and L1/L2 status. As shown in the previous literature, the quantified linguistic variance confirmed that there are more different ways to answer incorrectly than to answer correctly. Regarding gender, it was somewhat unexpected that boys exhibit statistically larger STTR values than girls, although the magnitude is very small at .011 when controlling for other variables. Similarly, it was unexpected that L1L2 status had no significant effect, although the direction of the estimate is consistent with what was observed in Figure 7 (right). With this smaller set of student characteristics as fixed effects, 27.1% of the linguistic variance was explained. Together with the random effects by level-2 and level-3 clusters, the total explained variance of STTR values was up to 84.9%. This is because student characteristics as level 1 predictors explained a large amount of the residual variance that was not modeled in M1 and M2.

Finally, when both item-related and student-related predictors were specified in M4, the variance explained by the fixed effects was about 51.1%. The pattern of significant variables remained the same, except for maximum score points and passage [T]. The significance level of Passage [T] was at the border in M2 ($p = .068$), so the effect of this variable seems to be sensitive to the model specifications. Given that there are relatively fewer items with maximum scores of 2 and 3, the result of this variable seems to be sensitive to other controlling variables as well.

Table 4

Effects of Item and Student Characteristics on Linguistic Variance

	Model 1		**Model 2**		**Model 3**		**Model 4**	
Predictors	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
(Intercept)	.573***	.012	.430***	.021	.646	.009	.512***	.021
Process [F]			–.082***	.011			–.069***	.012
Process [M]			–.001	.009			.012	.010
Process [I]			.038***	.009			.044***	.010
Passage [M]			–.003	.008			.000	.008
Passage [R]			.006	.009			.004	.010
Passage [T]			.017	.009			.028**	.010
Passage [Z]			.045***	.008			.050***	.009
Position			.004***	.001			.004***	.001
Max. Points [2]			.008	.007			.039***	.008
Max. Points [3]			–.037**	.014			.065***	.016
Difficulty			.239***	.020			.154***	.022
Gender [boys]					.011***	.002	.010***	.002
L1L2 [L2]					–.001	.003	–.001	.003
score [1]					–.140***	.002	–.138***	.002
score [2]					–.199***	.004	–.202***	.004
score [3]					–.316***	.008	–.322***	.008
Random Effects								
$\psi^{(2)}$	.012		.004		.014		.007	
$\psi^{(3)}$	.002		.003		.001		.002	
θ	.011		.011		.004		.004	
Marginal R^2	.000		.290		.271		.511	
Conditional R^2	.575		.553		.849		.837	

*** for $p < .001$, ** for $p < .01$, and * for $p < .05$

Relationship between Linguistic Variance and Scoring Performances

In this study, it is hypothesized that the linguistic variance impacts the performances of two automatic scoring approaches. First, the relationship between linguistic variance measured via STTR and the accuracy of the supervised classifiers based on semantics was examined (Figure 9 top left). Pearson correlation coefficient was estimated as –.235 ($p - value < .001$) and Spearman's rank correlation was –.314, both suggesting that less linguistic variance is associated with higher accuracy of the automatic scoring models. A similar analysis was also conducted for the accuracy of the fuzzy lexical matching (Figure 9 top right). The magnitude of correlation was not different from zero: Pearson's correlation coefficient was .004 ($p - value = .892$) and Spearman's rank correlation was -.028.

For the fuzzy lexical matching, the major performance metric is efficiency (i.e., the proportion of automatically scored responses)[6]. When efficiency was examined in relation to the STTR (Figure 9 bottom right), Pearson correlation coefficient was estimated as −.393 ($p - value < .001$) and Spearman's rank correlation was −.230, both again suggesting that less linguistic variance can lead to scoring more of students' text responses based on the fuzzy lexical matching approach.

To further investigate the relationship between linguistic variance and scoring performances, similar multilevel analyses were performed. Three types of dependent variables were used: 1) accuracy of supervised classifiers based on semantics approach (Accuracy (XLM-R)), 2) accuracy of fuzzy lexical matching approach (Accuracy (FLM)), and 3) efficiency of fuzzy lexical matching approach (Efficiency (FLM)). M1 (Equation 9) simply decomposes the variation of accuracy by language-country groups, M2 (Equation 10) is a simple random-intercept model with a $STTR_{ig}^{(1)}$ as a single predictor. M3 (Equation 11) was also performed to examine the effect of $STTR_{ig}^{(1)}$, after controlling for other item characteristics. It total, three multilevel models were used for each of the dependent variable. Model formulations are written as follows with the example of the accuracy of supervised classifier approach ($Accuracy^{(XLM-R)}$), and the same models were conducted with respect to the accuracy of fuzzy lexical matching ($Accuracy^{(FLM)}$) and the efficiency of the latter ($EFF^{(FLM)}$).

$$\begin{aligned} Accuracy_{ig}^{(XLM-R)} &= \beta + \zeta_g + \epsilon_{ig}, \\ \zeta_g &\sim \mathcal{N}(0, \psi), \\ \epsilon_{ig} &\sim \mathcal{N}(0, \theta). \end{aligned} \tag{9}$$

$$\begin{aligned} Accuracy_{ig}^{(XLM-R)} &= \beta_1 + \zeta_g + \beta_2 STTR_{ig}^{(1)} + \epsilon_{ig}, \\ \zeta_g &\sim \mathcal{N}(0, \psi), \\ \epsilon_{ig} &\sim \mathcal{N}(0, \theta). \end{aligned} \tag{10}$$

[6] Note that efficiency measure is not relevant for the supervised classifier approach since all text responses are automatically scored without exceptions, which means 100% efficiency.

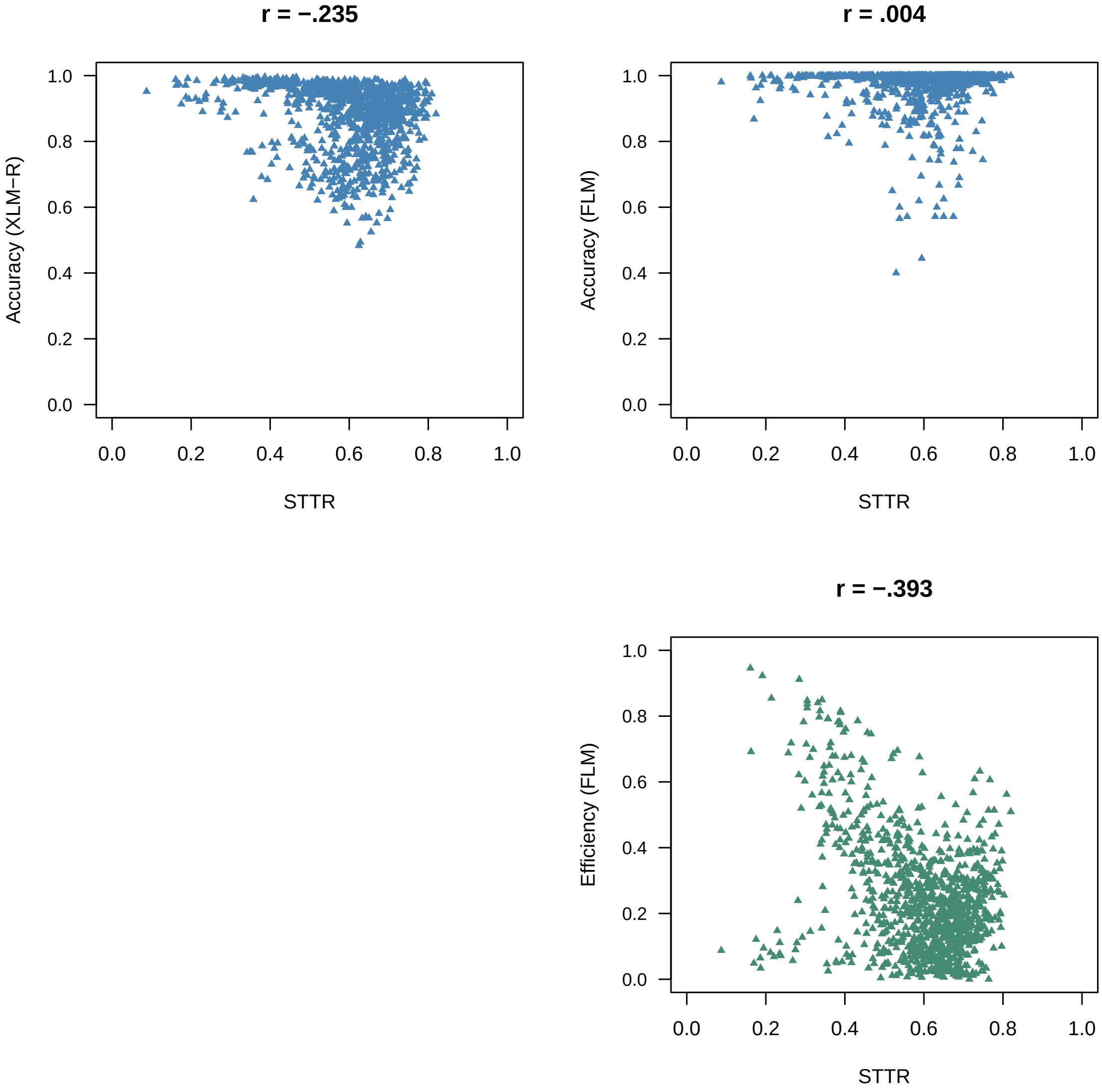


Figure 9. (a) Relationship between Accuracy (XLM-R) and STTR (top left) (b) Relationship between Accuracy (FLM) and STTR (top right), and (c) Relationship between Efficiency (FLM) and STTR (bottom right)

Table 5

Effects of Linguistic Variance and Item Characteristics on the Accuracy of XLM-R

	Model 1		**Model 2**		**Model 3**	
Predictors	Estimate	SE	Estimate	SE	Estimate	SE
(Intercept)	.877***	.005	1.036***	.018	.978***	.015
STTR			–.267***	.027	–.111***	.022
Process [F]					.032***	.007
Process [M]					.021***	.005
Process [I]					.020***	.005
Passage [M]					.011*	.005
Passage [R]					–.029***	.005
Passage [T]					.004	.005
Passage [Z]					–.011*	.005
Position					.000	.000
Max. Points [2]					–.166***	.005
Max. Points [3]					–.236***	.008
Difficulty					.007	.014
Random Effects						
ψ	$< .001$		$< .001$		$< .001$	
θ	.010		.009		.002	
Marginal R^2	.000		.099		.727	
Conditional R^2	.035		.182		.791	

*** for $p < .001$, ** for $p < .01$, and * for $p < .05$

$$\begin{aligned} Accuracy_{ig}^{(XLM-R)} &= \beta_1 + \zeta_g + \beta_2 STTR_{ig}^{(1)} + \beta_3 x_{3ig} + ... + \beta_k x_{kig} + \epsilon_{ig}, \\ \zeta_g &\sim \mathcal{N}(0, \psi), \\ \epsilon_{ig} &\sim \mathcal{N}(0, \theta). \end{aligned} \tag{11}$$

Accuracy of Automatic Scoring with XLM-R. Table 5 summarizes the results of three multilevel models with $Accuracy_{ig}^{(XLM-R)}$ as the dependent variable. M1 estimates the grand mean of accuracy as .877. While language-country membership explains 19.9% of the total variance in linguistic variance (STTR), it only explains approximately 3.5% of the variation in accuracy. This finding is highly relevant because it suggests that the multilingual automatic scoring approach used in this study can be applied to different languages and countries, despite their linguistic differences reflecting their unique language-country characteristics.

Table 6

Effects of Linguistic Variance and Item Characteristics on the Efficiency of FLM

	Model 1		**Model 2**		**Model 3**	
Predictors	Estimate	SE	Estimate	SE	Estimate	SE
(Intercept)	.246***	.012	.703***	.031	.579***	.041
STTR			−.769***	.041	−.626***	.060
Process [F]					.106***	.019
Process [M]					.020	.015
Process [I]					−.018	.015
Passage [M]					−.102***	.012
Passage [R]					−.047**	.014
Passage [T]					−.060***	.015
Passage [Z]					−.086***	.014
Position					.004***	.001
Max. Points [2]					−.080***	.012
Max. Points [3]					−.079***	.022
Difficulty					.122**	.037
Random Effects						
ψ	.002		.006		.005	
θ	.030		.022		.017	
Marginal R^2	.000		.250		.379	
Conditional R^2	.073		.416		.527	

*** for $p < .001$, ** for $p < .01$, and * for $p < .05$

Efficiency of Fuzzy Lexical Matching. Finally, Table 6 summarizes the results of the efficiency measure for the fuzzy lexical matching approach with $EFF_{ig}^{(FLM)}$ as the dependent variable. Since fuzzy lexical matching only provides automatic scores based on a stochastic matching algorithm, only 24.6% of text responses could be scored by the machine, while approximately 75% of responses still required human scoring. Language-country groups can explain less of the variation in efficiency, about 7.3%, compared to the 19.9% explained in the linguistic variance (STTR), which empirically supports the largely language-agnostic feature of the FLM system. The matching algorithm becomes more powerful when text responses are more homogeneous. This intuition is also confirmed by the significant relationship with STTR shown in M2: a 0.10 increase in STTR value is likely to increase human scoring work by as much as 7.7%. After controlling for item characteristics, the decrease in efficiency associated with a 0.10 increase in STTR value reduced to 6.3%, as shown in M3.

Table 7

Effects of Linguistic Variance and Item Characteristics on the Accuracy of FLM

	Model 1		**Model 2**		**Model 3**	
Predictors	Estimate	SE	Estimate	SE	Estimate	SE
(Intercept)	.971***	.003	.986***	.011	.971***	.015
STTR			–.026	.017	.032	.023
Process [F]					.004	.008
Process [M]					.003	.006
Process [I]					–.004	.006
Passage [M]					–.001	.005
Passage [R]					–.007	.006
Passage [T]					–.021***	.006
Passage [Z]					–.014*	.006
Position					.000	.000
Max. Points [2]					–.070***	.005
Max. Points [3]					–.069***	.009
Difficulty					.022	.015
Random Effects						
ψ	<.001		<.001		<.001	
θ	.004		.004		.003	
Marginal R^2	.000		.011		.608	
Conditional R^2	.037		.047		.307	

*** for $p < .001$, ** for $p < .01$, and * for $p < .05$

Accuracy of Fuzzy Lexical Matching. The fuzzy lexical matching system produced a grand mean accuracy of .971. This high level of accuracy is not unexpected, as the MSCS version of the FLM system was designed to achieve 100% accuracy at the expense of efficiency. Language-country membership explains only 3.7% of the variation in accuracy with FLM. This suggests that a FLM system can be used for different languages and countries, regardless of their linguistic variation, similar to the approach building on XLM-R. The matching procedure already takes into account linguistic variance when the text response receives an automatic score. Therefore, it is not surprising that the effect of linguistic variance is not significant in both M2 and M3. The linguistic variance explains only about 1% of the accuracy of the FLM system.

Discussion

Conclusion

The aim of the current project was (1) to analyze the performance of automatic scoring of text responses in the 2016 ePIRLS, (2) to assess the linguistic variance in text responses as a pivotal determinant of automatic scoring performance, and (3) to investigate the impacts of student and item characteristics on the linguistic variance, and in turn, automatic scoring performance. For automatic scoring, we used two systems: fuzzy lexical matching (FLM) and supervised classifiers based on semantics (XLM-R). Fuzzy lexical matching prioritizes the accuracy (in theory, 100%), and a stricter version (MSCS) has been operationalized in the PISA since the 2018 cycle. However, this system cannot score all text responses. As effective and simple it is for items that elicit a high degree of regularity in responses, it is equally limited in its application to items with medium and low levels of regularity. Therefore, we used efficiency (i.e., the proportion of text responses that are scored by machine) as a major metric to evaluate the performance of this system, indicating the potential for manual effort reduction. The second system, supervised classifiers based on semantics, extracted semantic representational features by employing a pre-trained multilingual deep neural network (XLM-R), followed by model training with support vector machines. With this approach, all text responses are scored with varying accuracy. Thus, performance of this system was evaluated regarding agreement with human raters (i.e., quadratic weighted kappa and exact agreement rates). In the process of automatic scoring, we view linguistic variance as a pivotal determinant that impacts the performance of automatic scoring systems, along with other instrument-related and student-related factors. Therefore, we computed a modified type-token ratio, $STTR$, as a measure of linguistic variance to enable comparison across groups of students with similar characteristics such as gender, L1/L2 speaker status, and item-level score, taking into account items and language-country groups. We then analyzed corresponding relationships using multiple multilevel models. The results can be summarized as follows.

First, the accuracy of the automatic scoring with XLM-R was satisfactory, with an average agreement across all items and language groups of $\kappa = .755$. Out of a total of 50 items, 36 show an agreement of at least $\kappa \geq .650$, 22 items of at least $\kappa \geq .750$, and 11 items of at least $\kappa \geq .850$. Automatic scoring with FLM showed a decent performance in efficiency, ranging from 17.1% to 37.85%, with a median of 26.1%. This suggests that approximately 75% of text responses still require scoring work from humans but also that about 25% of manual scoring effort could be saved. Although not directly comparable, the reduction of scoring work in the 2016 ePIRLS is slightly higher than that of the operationalized system (which is a stricter version of FLM) in PISA 2022, where the median was reported to be 21.8% for the reading domain (OECD, 2023).

Second, when the linguistic variance was computed at the item level for each language-country group ($STTR^{(1)}_{ig}$), 19.9% of linguistic variance was explained by language-country groups. When computing linguistic variance at a more granular level based on student characteristics ($STTR^{(2)}_{pig}$), the total variance was decomposed into 9.3% of between-language-country groups and 48.2% of between-item, within-language-country groups. Although not detailed in the paper, an alternative cross-classified structure was attempted (Shin, Rabe-Hesketh, & Wilson, 2019), where students nested in items and nested in countries, while items were crossed with countries. The results provided a similar story, showing 9.0% by language-country-specific factors, while 61.1% by item-specific factors.

Third, instrument-related characteristics considered in this study were mostly significant to the linguistic variance. They include the cognitive process being measured by an item, the passage to which it belongs, the position of the item within the passage, the maximum score points, and the item's difficulty. When combined, these characteristics could explain 50.1% of item-level linguistic variance ($STTR^{(1)}_{ig}$) and 29.0% of student-level linguistic variance ($STTR^{(2)}_{pig}$). Among student-related characteristics, gender, and the item-level score were statistically significant, while L1/L2 status did not have a significant effect. In the most general model that specified both instrument-related and student-related characteristics (M4 in Table 4), 51.1% of linguistic variance was explained. When combined with random effects from items and language-country groups, 83.7% of the total linguistic variance was explained.

Fourth, importantly, language-country groups were only marginally related to the performance of automatic scoring; 3.5% for the XLM-R's accuracy and 7.3% of FLM's efficiency.[7] In contrast, it should be noted that the corresponding values for explaining the item-level linguistic variance ($STTR^{(1)}_{ig}$) were 19.9% and 48.2%, respectively, for student-level linguistic variance ($STTR^{(2)}_{pig}$). These quantities confirm the potential to use automatic scoring models accurately and efficiently in a multilingual context across countries, regardless of their linguistic variances.

Lastly, as hypothesized in the study, linguistic variance exhibited significant negative impacts to the performances of automatic scoring. The increase of linguistic variance was significantly associated with the decrease of accuracy of XLM-R (Table 5) and the efficiency of FLM (Table 6). In contrast, because FLM took into account linguistic variance in the matching procedure, linguistic variance did not impact the accuracy of FLM (Table 7).

Implications for Operational Use

The feasibility of operational automatic scoring remains a challenge for large-scale international assessments such as PIRLS. Barriers to automatic scoring include the multilingual nature of text responses in international studies, coupled with the requirement for language-dependent natural language processing resources, and the effort required to develop, quality control, and maintain automatic scoring models that meet satisfactory levels of accuracy. Moreover, the dynamic developments of large-scale assessments, such as their frameworks and new items, can hamper its re-usability. Given the high stakes of PIRLS assessment results at the policy level, the operationalization of the automatic scoring system requires a careful approach that can ensure the reliability, validity, and comparability of results across participating countries and cycles. In this study, we attempted two complementary automatic scoring approaches; one that prioritizes accuracy at the expense of not scoring some text responses by machine (FLM) and another that utilizes state-of-the-art multilingual natural language processing technologies. In the light of the result highlights elaborated on above, we conclude the following implications for the operational use of automatic scoring.

First, two automatic scoring models, FLM and XLM-R, empirically demonstrated their potential for automatic scoring of the 2016 ePIRLS data in terms of efficiency and ac-

[7] The corresponding value for the accuracy of the FLM was 3.7%, which is close to the 3.5% of the XLM-R approach.

curacy. The amount of saved human resources by implementing a simple system, such as FLM, would substantially reduce costs and required time for scoring, and, moreover, improve scoring quality. Because the two approaches were designed to be largely language-agnostic (FLM) and multilingual (XLM-R), the extent to which the language-country group explained the performance of the automatic scoring models was very small. In contrast to the ability of language-country groups to explain linguistic variation, such results made it clear that these two automatic scoring models can be used independently of linguistic variation for the examined test languages, which covered a wide range of language families and scripts.

Second, an ensemble framework that implements FLM and XLM-R in a sequential manner can be considered for operationalization. To achieve the highest level of reliability, validity, and comparability of automatically scored responses, FLM could first be implemented on the entire set of incoming unscored responses for trend or linking items. The remaining, unmatched text responses written in different languages, both for trend/linking and new items, could then be scored by a system such as XLM-R. The current evaluation of XLM-R reported in this study has not yet implemented fine-tuning, but has included such noisy data that it contains a substantial amount of misclassifications. Thus, if there is operational room to develop pipelines for fine-tuning XLM-R system and improving the quality of training data, the performance of XLM-R system can be greatly improved.

Third, as hypothesized in our study according to Horbach and Zesch (2019), linguistic variance turned out to be a pivotal factor affecting the performance of automatic scoring. Different components of linguistic variances were shown to be affected by certain student- and, more importantly, instrument-related characteristics. This allows a-priori judgments which items may be suited for automatic scoring and which ones should remain in the manual scoring process as the development of accurate classifiers might currently not be possible yet.

Fourth, since the study revealed a significant number of inconsistent human scores, there is an opportunity for two additional improvements. First, the operational use of automatic scoring could improve scoring quality by highlighting inconsistencies within and between human raters and reducing the number of text responses that need to be scored manually.

Similarly, calibration sessions at the beginning of scoring sessions could be improved by selecting critical examples from the empirical data. On the other hand, both automatic scoring systems could help to identify and correct gaps in the scoring guides, if common responses are not covered at all as reference texts, or if very similar responses are coded inconsistently, which might indicate unclear scoring guides.

Limitations

The present study revealed novel findings but has a number of limitations. First, we restricted the multilevel model specifications to main effects for ease of interpretation. However, more complicated relationships may be of interest, including interaction terms between student-related, item-related, or cross-student item-related characteristics. In addition, random slopes may be considered if the effects of some item or student characteristics are expected to vary across language-country groups.

Currently, since our main focus was to decompose the variance by different clusters, and there was no software package available for fitting multilevel beta regression models, we decided to treat the outcome variable as a continuous variable, which allowed for multilevel analyses. More fundamentally, given that all outcome variables are proportions between 0 and 1, a beta regression would have been a better methodological choice. Alternatively, with respect to the distribution of the outcome variable, zero-inflated and hurdle models (Lambert, 1992) may be more appropriate, using the flexible Bayesian approach provided in the `Rstan` package (Stan Development Team, 2023).

Next, we conducted a limited examination of the effects of student characteristics. In particular, students' overall performance score or proficiency level was not included in our analyses. This is because including this variable in addition to gender, L1/L2 speaker status, and item-level score made the calculation of $STTR_{pig}^{(2)}$ extremely sparse. In addition, student-related characteristics could not be specified when examining the effects of linguistic variance on automatic scoring performance (see relevant Tables 5, 6, and 7). All types of three-level models, including variance-components model, were attempted, but all failed to converge. Therefore, student-related factors were not controlled for in modeling the relationship between linguistic variance and automatic scoring performance. To this end, more sophisticated modeling strategies can be explored that would ultimately help address issues of fairness in automatic scoring.

The texts were processed in their raw form for the XLM-R analyses, without any preprocessing such as decapitalization or spelling error correction. Due to the flexible tokenization of the RoBERTa language model, no information was lost in this process. However, text responses may then be treated differently, especially if the texts contain deviating spellings, such as misspellings or uppercase letters only, leading to divergent semantic vector representations. Given the high frequency of misspellings due to the low age of the target population, it cannot be ruled out that preprocessing, especially spelling correction, could increase the reliability of automatic scoring of some items and, thus, improve the scoring quality.

Scoring performance was assessed regarding the alignment of automatic and manual scores. Since responses can be scored differently by two raters due to distinct interpretations, further inter-rater reliability measures are essential to compare machine-human agreement with human-to-human agreement. Responses poorly coded by machines, potentially due to complex linguistic structures, might also be inconsistently scored by human raters, a situation likely in cases of polytomously scored responses. Hence, an inter-rater reliability study might provide a benchmark of human scoring for interpreting the performance of corresponding automatic scoring.

Moreover, STTR is only one potential operationalization of linguistic variance of many and focuses on relative lexical diversity. In follow-up analyses, we will consider other operationalizations such as pairwise distances and semantic variance. Finally, the significant number of human misclassifications, among others apparent in the large number of empty responses with positive scores or deviating scores for identical responses, hampered the training and evaluation of both automatic scoring systems to a certain extent.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. Retrieved from http://epm.sagepub.com/content/20/1/37.short doi: 10.1177/001316446002000104

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., … Stoyanov, V. (2020, July). Unsupervised cross-lingual representation learning at scale. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 8440–8451). Online: Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747

Fleiss, J. L. (1981). The measurement of interrater agreement. In J. L. Fleiss, B. Levin, & M. C. Paik (Eds.), *Statistical methods for rates and proportions* (pp. 598–626). New York: John Wiley & Sons.

Haller, S., Aldea, A., Seifert, C., & Strisciuglio, N. (2022). *Survey on automated short answer grading with deep learning: from word embeddings to transformers.*

Horbach, A., & Zesch, T. (2019). The Influence of Variance in Learner Answers on Automatic Content Scoring. *Frontiers in Education*, *4*, 4. doi: 10.3389/feduc.2019.00028

Koizumi, R. (2012). Relationships between text length and lexical diversity measures: Can we use short texts of less than 100 tokens. *Vocabulary Learning and Instruction*, *1*(1), 60–69.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, *34*(1), 1–14.

Martin, M. O., Mullis, I. V., & Foy, P. (2016). Assessment design for pirls, pirls literacy, and epirls in 2016. *PIRLS*, 55–69.

Mullis, I. V., & Martin, M. O. (2019). *Pirls 2021 assessment frameworks.* ERIC.

Mullis, I. V. S., & Martin, M. O. (2015). *PIRLS 2016 assessment framework* (2nd ed.). International Association for the Evaluation of Educational Achievement.

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2017). *ePIRLS 2016: International results in online informational reading.* International Association for the Evaluation of Educational Achievement.

OECD. (2019). *PISA 2018 Technical Report: Coding design, coding process, coding reliabil-*

ity studies, and machine-supported coding in the main survey (Chapter 13). Retrieved 2020-09-21, from https://www.oecd.org/pisa/data/pisa2018technicalreport/PISA2018-TecReport-Ch-13-Coding-Reliability.pdf

OECD. (2023). *PISA 2022 Technical Report: Coding Design, Coding Process, and Reliability Studies (Chapter 15).* Retrieved 2023-12-18, from https://www.oecd.org/pisa/data/pisa2022technicalreport/PISA-2022-Technical-Report-Ch-15-PISA-Coding-Reliability.pdf

Shin, H. J., Rabe-Hesketh, S., & Wilson, M. (2019). Trifactor models for multiple-ratings data. *Multivariate Behavioral Research*, *54*(3), 360–381.

Stan Development Team. (2023). *RStan: the R interface to Stan.* Retrieved from https://mc-stan.org/ (R package version 2.32.3)

Sukkarieh, J. Z., Von Davier, M., & Yamamoto, K. (2012). From biology to education: Scoring and clustering multilingual text sequences and other sequential tasks. *ETS Research Report Series*, *2012*(2), i–43.

Whitmer, J., Deng, E., Blankenship, C., Beiting-Parrish, M., Zhang, T., & Bailey, P. (2023). *Results of NAEP Reading Item Automated Scoring Data Challenge (fall 2021).* EdArXiv. Retrieved from https://edarxiv.org/2hevq doi: 10.35542/osf.io/2hevq

Yamamoto, K., He, Q., Shin, H. J., & von Davier, M. (2018). Development and implementation of a machine-supported coding system for constructed-response items in PISA. *Psychological Test and Assessment Modeling*, *60*(2), 145–164.

Yamamoto, K., He, Q., Shin, H. J., & von Davier, M. (2017). Developing a Machine-Supported Coding System for Constructed-Response Items in PISA. *ETS Research Report Series*, *2017*(1), 1–15. doi: 10.1002/ets2.12169

Yaneva, V., & von Davier, M. (2023). *Advancing natural language processing in educational assessment.* Taylor & Francis.

Zehner, F., Sälzer, C., & Goldhammer, F. (2016). Automatic coding of short text responses via clustering in educational assessment. *Educational and Psychological Measurement*, *76*(2), 280–303. doi: 10.1177/0013164415590022

Zehner, F., Shin, H. J., Kerzabi, E., Andersen, N., Goldhammer, F., & Yamamoto, K. (2021). Automatic normalization for advancing response coding consistency and efficiency in PISA. In National Council on Measurement in Education (Ed.), *Proceedings of the NCME 2021 Annual Meeting.* Baltimore, MD.

Zesch, T., Horbach, A., & Zehner, F. (2023). To score or not to score: Factors influencing performance and feasibility of automatic content scoring of text tesponses. *Educational Measurement: Issues and Practice.* doi: 10.1111/emip.12544

Acknowledgements

We extend our sincere thanks to the IEA for funding this study. Special thanks go to Lauren Musu, Ralph Carstens, and Mark Cockle for their support in this research and their prompt attention to our requests. Data management for a study as large as PIRLS is challenging enough, but handling the labyrinthine world of text responses in a variety of scripts is even more so. We greatly appreciate their efforts to ensure that we could work with the highest quality data possible.