Did the Names I Used within My Essay Affect My Score? Diagnosing Name Biases in Automated Essay Scoring

Ricardo Muñoz Sánchez[†], **Simon Dobnik**[‡], **Maria Irena Szawerna**[†], **Therese Lindström Tiedemann**[§], **Elena Volodina**[†]
[†] Språkbanken Text, University of Gothenburg, Sweden
[‡] CLASP, FLoV, University of Gothenburg, Sweden
[§] Department of Finnish, Finnougrian and Scandinavian Studies, University of Helsinki, Finland
mormor.karl@svenska.gu.se
[†] [‡]{ricardo.munoz.sanchez,simon.dobnik,maria.szawerna,elena.volodina}@gu.se
[§]therese.lindstromtiedemann@helsinki.fi

Abstract

Automated essay scoring (AES) of second-language learner essays is a high-stakes task as it can affect the job and educational opportunities a student may have access to. Thus, it becomes imperative to make sure that the essays are graded based on the students' language proficiency as opposed to other reasons, such as personal names used in the text of the essay. Moreover, most of the research data for AES tends to contain personal identifiable information. Because of that, pseudonymization becomes an important tool to make sure that this data can be freely shared. Thus, our systems should not grade students based on which given names were used in the text of the essay, both for fairness and for privacy reasons. In this paper we explore how given names affect the CEFR level classification of essays of second language learners of Swedish. We use essays containing just one personal name and substitute it for names from lists of given names from four different ethnic origins, namely Swedish, Finnish, Anglo-American, and Arabic. We find that changing the names within the essays has no apparent effect on the classification task, regardless of whether a feature-based or a transformer-based model is used.

1 Introduction

Artificial intelligence is being deployed in high-stakes situations, such as automated grading of second language essays in proficiency assessment. While AI can improve the opportunities students have in education, the job market, etc., such systems often display human-like biases (Blodgett et al., 2020). Aldrin (2017) notes that human graders have a slight bias based on names appearing in essay texts. In this paper we aim to identify whether the same pattern holds in automated systems.

The broad question for our study is: are there any implicit biases that models have learnt from the training data that can influence automated essay scoring in a negative way? In particular, we are interested in uncovering potential biases that can be associated with use of names representing different ethnic groups – and how this can be reflected in the domain of automatic essay scoring (AES).

For the purposes of this work, we say that there is bias in AES when an essay is scored not only by its contents but also by the assumed demographic characteristics of its author. We use this definition as we are looking for biases in a downstream application (i.e. extrinsic biases) as opposed to biases either in the training data or in any intermediate representations (i.e. intrinsic biases). Even though we know that biases in deep learning models cannot be removed in absolute terms (Gonen and Goldberg, 2019), we can attempt to minimize their impact.

Because of this, we have set out to create a novel paradigm of diagnostic benchmarks for identifying hidden biases in AES models as a safety gatekeeping before they are approved for use in real-life scenarios. In such a dataset each essay is duplicated (several times), artificially altering given names appearing in the text to identify if such perturbation affects how an essay is scored. Since the essays are identical as far as linguistics, language complexity, and content are concerned, we expect them to be graded similarly. Thus, we would say that our model for this task presents bias if it systemically assigns lower grades when using versions of the essays with names coming from specific ethnicities.

Our research questions are the following:

- Does changing given names inside a second language learner essay affect the way the text is graded when using automated essay scoring?
- How much does this differ between feature-based machine learning and deep learning?

For this, we use a de-anonymized (i.e. origi-

Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization
CALD-pseudo 2024, pages 81–91

nal) version of the SweLL-pilot corpus of second language Swedish learner essays (Volodina et al., 2016a), which consists of 502 essays annotated with CEFR levels[1] (Council of Europe, 2001), as our source data.

First, we compile four lists of given names inspired by those in Aldrin (2017): traditional Swedish names; modern Swedish names of Anglo-American origin; Finnish names (due both to the close sociocultural links between Finland and Sweden and to Swedish being an official language of Finland being learnt by the population that does not speak it as their first language); and names of Arabic origin (the most prominent group of learners in the corpus).

Second, we create a diagnostic dataset to identify biases in the classification task. We select SweLL-pilot essays in which a given name appears only once. Then, we generate an essay version for each name on the lists by substituting the name in the original text with one from the list. All of the essays chosen have the names in their base form.

Third, we fine-tune a BERT (Devlin et al., 2019) model on the original SweLL-pilot data to predict the CEFR level of a given essay and compare it to an existing feature-based model (Pilán et al., 2016).

Finally, we test the two models and compare the equality of opportunity between the different given name groups on the diagnostic dataset, as described by Hardt et al. (2016).

As mentioned previously, we would expect an unbiased or a fair model (in terms of given names) to not show systemic misclassification for the ethnic groups considered. It does not mean that it will be unbiased towards names from other ethnic groups or that different names would not elicit unexpected responses from our model (Antoniak and Mimno, 2021). It is important to note that a model being fair for a downstream application does not mean that the model itself, the data, or the annotation lack biases (and vice versa). Social biases are a very complex phenomenon and they can be embedded in a variety of ways, as illustrated by Suresh and Guttag (2021). Moreover, Goldfarb-Tarrant et al. (2023) note that the presence or absence or intrinsic biases (e.g. in language models) does not necessarily correlate with the presence or absence of extrinsic biases (e.g. in downstream applications). Because of this, it is important to monitor and to audit AES models regularly regardless of whether they are fair. And, given that this is a high-stakes task, it is essential to always have a human-in-the-loop approach.

The rest of the paper is structured as follows: Section 2 reviews some of the related work both in terms of automated essay assessment and of bias and fairness in NLP. Section 3 presents our methodology, the models and data we used, as well as how they were evaluated. In Section 4 we show and discuss the results from our experiments, while in Section 5 we present some ideas for future work.

2 Related Work

Language assessment and subsequent documented language proficiency, be it for citizenship, university admission or a job application, are extremely influential, if not life-changing, both on the individual, societal and political levels (Roever and McNamara, 2006). Assessment should therefore be guaranteed to be fair and unbiased, and assessors should be kept accountable for the results, i.e. be able to motivate the assigned scores (e.g. ASLHA, 2023; ALTE, 2020). This is a non-trivial requirement even for human assessors, and is clearly a much greater challenge for automated language assessment.

2.1 Biases in Humans

People carry a multitude of implicit associations which have been acquired through previous experiences, for example, an association between '`day and ...`' (`night`, supposedly) or '`commit a ...`' (`crime`, most probably). These associations are called *implicit biases*, which can be neutral, positive or negative in nature. Implicit associations (or biases) do not necessarily have an impact on the life around us, but in certain cases they do – and then they can risk jeopardizing our ideals of fairness and equality, for example when it comes to racial or gender discrimination (Greenwald et al., 2015). Especially important are the associations that are triggered in ambiguous and confusing contexts, when our brain falls back on the associations stored in our memory from earlier experiences, especially those that are stored repeatedly (Greenwald and Krieger, 2006).

For example, Foster (2008) has suggested that there may be a correlation between ethnicity and (lower) results at a university, but that this is un-

[1] CEFR stands for Common European Framework of Reference for Languages. It is a framework to evaluate foreign language learning and assigns one of six reference levels to determine the proficiency level of a second language speaker.

likely to be directly due to ethnically marked names at that stage of education. Aldrin (2017) took it further and investigated whether there was an influence from stereotypically marked given names in first language Swedish essays by letting 113 human assessors mark one text where she inserted, in quite a discrete place, one of three names: a traditional Swedish name, an ethnically marked name or an Anglo-American name with certain socio-economical associations. The results showed a certain influence on the assessment of language proficiency, "stylistic precision" and "writing technique", but nothing statistically significant. She believed the fact that the results were not as clear as in previous international work (e.g. Anderson-Clark et al., 2008; Figlio, 2003) could be due to the fact that (1) the name was discreetly placed, and (2) several of the teachers worked in schools with students of heterogeneous background and were therefore less likely to have a bias, or (3) that the names picked were not found to be stereotypical in the way they were thought to be.

2.2 Biases in Machine Learning Systems

Similar to humans, language models, including Large Language Models (LLMs), store associations between various linguistic and non-linguistic information types that they meet during the training stage. These models do this by looking at large amounts of data, finding patterns and repeating them. An important issue here is that social biases are also reflected in the data that we, humans, produce (Marchiori Manerba et al., 2022), leading to models that parrot sexist (e.g. Zhao et al., 2018), racist (e.g. Sap et al., 2019), or xenophobic (e.g. Narayanan Venkit et al., 2023) ideas.

Following Blodgett et al. (2020), we claim that any work on biases in NLP and AI-based systems should be well-grounded in the domain where biases need to be uncovered, since (negative) biases in one domain are not necessarily negative in another. For example, absence of Past Simple in an essay is an indication of a lower grade. However, it might not be a negative feature when applied to filtering application letters for appropriate job candidates. Therefore studying biases in a vacuum can be misleading for a particular domain.

Some previous papers have studied biases regarding names in NLP. Several of the word embedding association tests (WEAT, Caliskan et al., 2017) compare lists of Anglo-American and Afro-American given names and lists of stereotypical characteristics associated with each. Meanwhile, several studies have found that the appearance of names in text can affect how it is translated (e.g. Wang et al., 2022; Sandoval et al., 2023). Furthermore, some studies have seen how nationalities and names of countries are related to the text that auto-decoders generate (e.g. Narayanan Venkit et al., 2023).

2.3 Biases in Automated Essay Scoring

Concerns about risks of introducing biases into automatic assessment scores have also been raised. Some studies criticize automatic essay scoring algorithms for flawed grading of high-stakes exams pointing out bias against certain demographic groups[2] (e.g. Madnani et al., 2017; Loukina et al., 2019) due to data imbalance or rater bias reflected in the data. Despite the criticism, the technology has been embraced and has shaped life stories of thousands of people.

Kane (2001) views validity and fairness in language assessment as closely related ways of looking at the same question. That is, whether the proposed interpretations and uses of test scores are appropriate for a population over some range of contexts. The traditional definition of fairness in the field of educational measurement is when a test does not unduly advantage or disadvantage any groups (Kane, 2001). The concept of fairness is also closely connected to bias, or the lack thereof. Bias is when the validity of a given test score is different for subgroups of test-takers. For example, this may happen if a set of items would favor a particular group in a given test. Test scores would then not reflect the participants' true ability.

To overcome the technological biases, Madnani et al. (2017) suggest a scheme to detect demographic and construct-irrelevant biases (such as rater biases, data-imbalance, machine-learning biases) applying model validation based on psychometric and statistical checks using an open-source tool RSMTool.[3] They also suggest reducing susceptibility to construct-irrelevant factors by design, among others by using feature review by experts and combining features into several models by feature type instead of mixing all features in one model. However, more advanced machine learning and neural network algorithms and LLMs are not

[2] https://www.vice.com/en/article/pa7dj9/flawed-algorithms-are-grading-millions-of-students-essays

[3] https://rsmtool.readthedocs.io/

as easily interpretable (Alishahi et al., 2020), which requires other approaches and solutions.

3 Experiment Setup, Materials and Methods

Our major question for the experiment is whether algorithms for essay classification are sensitive to names (or pseudonyms) used in essays. If we need to pseudonymize research data on a constant basis to protect writer identities, which is a GDPR requirement (EU Commission, 2016), we should find ways to do so that do not affect students. This means that it is our responsibility to check the effect the replacement candidates may have on the data and its downstream tasks and research applications. In this experiment we study the effects that replacing given names in learner essays might have on essay assessment in terms of CEFR[4] level, as described in the Introduction (section 1). The CEFR levels are a six-level scale to gauge the proficiency of an individual on a foreign language (i.e. not their first language or languages) and they range from A1 to C2, with A1 being the lowest.

3.1 Dataset

For our experiments we use SweLL-pilot (Volodina et al., 2016a; Volodina, 2024), a corpus of essays written by learners of Swedish as a second language (L2 Swedish). It contains 502 essays labeled with CEFR levels, distributed as shown in Table 1. Given the specifics of learner essays, many of them touch on personal stories, mostly in response to topics like 'The best day of my life', 'My school', 'My best friend', etc., which, of course, elicits a lot of private or sensitive information, starting with personal names, place names and other information that can reveal the writer's identity either in a direct or in an indirect way. This is natural, given that some of the CEFR levels expect the student to be able to describe topics about the personal lives.

To select essays for purposes of identifying biases based on given names, a few guidelines were applied:

- there should be, optimally, only one personal name used in its base form in each essay;
- if possible, no geographical context of the country of origin should be present;
- two essays per level are included.

[4]CEFR stands for Common European Framework of Reference for Languages.

Level	# essays	# of diagnostic essays	
		original	pseudonymized
A1	59	2	160
A2	143	2	160
B1	86	2	160
B2	105	1	80
C1	96	0	0
C2	7	0	0
Total	497	7	560

Table 1: Number of essays in the SweLL-pilot corpus per CEFR level, and statistics over the diagnostic dataset.

These guidelines aim to modify as little as possible in the text of the essays. This should allow for more controlled experimentation, leading in turn to a better way to ascertain the presence or absence of biases.

The selection proved to be more challenging than expected. First of all, in essays where personal information was elicited through a topic, usually more than one name were used, e.g. 'I have five brothers: name1, name2, name3, ...'. Second, the higher levels in the corpus (B2, C1 and C2) contain practically no essays where personal information is provided. This is due to the topics present in the dataset being of a non-personal nature at higher levels of proficiency, e.g. book reviews, argumentative essays and the like. We have, therefore, limited the diagnostic dataset to levels A1, A2, B1 and B2, with only one original essay for B2. No essays were found to meet our requirements at levels C1 and C2.

The IDs of the selected essays can be found in Appendix A and we call the resulting dataset with substituted names the diagnostic dataset.

3.2 Name Selection

The names used to check for biases were inspired by those chosen by (Aldrin, 2017). The idea behind this is to allow for better comparison in terms of the kinds of social biases we expect to find. In general, the idea is to compare how the model perceives stereotypical Swedish given names in the essays in comparison to those that are not usually associated with people with a Swedish background, particularly those that people in Sweden may be familiar with through their social contact.

We balance the different name lists by (binary)

gender[5] and by name group. Thus, we got 10 names for each combination of gender + name group, 20 names for each group and 80 names in total. The full lists of names can be found in Appendix B.

As mentioned in Section 1, we have chosen the following four name groups:

- Swedish names, taken from lists containing the top 100 given names normally used by men and by women[6]. These lists were obtained from Statistics Sweden, an official government website dedicated to publishing statistics about the country. This group was chosen as we are dealing with essays written in Swedish in Sweden. Furthermore, we made sure that none of the names chosen for the three originally non-Swedish given names appeared in other two lists.

- Finnish names, taken from lists of the top 10 first names throughout different decades[7]. This list was obtained from the Digital and Population Data Services Agency in Finland. This group was chosen due to Finland's and Sweden's close historical and cultural proximity and because Swedish is also one of the official languages in Finland, which means that it is not uncommon that students have to take exams in that language. As with all of the other groups other than the Swedish name, particular care was put into looking for names that are used as given names in Sweden, while checking that they do not overlap with common Swedish names.

- Anglo-American names, taken from the list of the top 100 names over the last 100 years in the United States[8]. This list was obtained from the Social Security Administration of the United States. This group was chosen as popular culture from the United States has permeated different countries in different ways. On top of that, these names can have different socio-cultural connotations in non-English speaking countries, including Sweden (Malm and Zetterström, 2007).

- Arabic names, taken from lists of commonly used Moroccan names used in the Netherlands (Gerritzen, 2007) and of commonly used Syrian names in Sweden (Gustafsson, 2021). These lists were later cross-referenced with information from Statistics Sweden[9] to verify that they are indeed commonly used given names in Sweden without being traditional Swedish names.

It is important to note that we combined different spellings of these names and kept just the one that is the most common in a Swedish context. This was necessary both to ensure that all of the lists contain the same amount of names and to keep the lists with as little overlap as possible (e.g. not including Sarah in the Anglo-American list as Sara was already in the Swedish list).

3.3 Models

We compare biases on the automated essay scoring task on two models, one feature-based and the other using a transformer architecture. The idea being that a feature-based system that does not explicitly use proper names should not exhibit name-based biases, while a model based on distributional semantics might pick up unwanted biases during its pre-training along all of the useful semantic information.

The feature-based approach we follow is that of Pilán et al. (2016) and Volodina et al. (2016b). They extract length-based, lexical, morphological, syntactic, and semantic features. Then they use an SVM as a classifier as well as feature selection and found that lexical features work best for classification. Even though they did not use any features that directly relate to proper names, there are some that are based on token length and some names that are also common nouns might appear in frequency-based lists (for example Hope in English).

The dataset used originally was SweLL-pilot (Volodina et al., 2016a) and they used adjacent accuracy to evaluate the model. What is, they treat the classes as an ordinal scale and consider that an answer was correct if it was either the correct class or the immediate one either before or after. That is under the intuition that misclassifying an A2 essay

[5] Finding common gender-neutral names proved to be a challenge as both the papers and the government agencies we consulted only listed male and female names.

[6] https://www.scb.se/en/finding-statistics/statistics-by-subject-area/population/general-statistics/name-statistics/

[7] https://verkkopalvelu.vrk.fi/nimipalvelu/default.asp?L=3

[8] https://www.ssa.gov/OACT/babynames/decades/century.html

[9] https://www.scb.se/en/finding-statistics/sverige-i-siffror/namesearch/

as B1 is a smaller mistake than misclassifying is as a B2 or C1 essay. Do note that we do not use this metric for this work, we report regular accuracy instead. This is, to the best of our knowledge, the current state of the art regarding CEFR level assessment in Swedish.

We also use a transformer-based model for our experiments to see whether their contextual behavior leads to biases in AES. This is a Swedish version of BERT trained by KBLab[10] (Malmsten et al., 2020), the NLP research group at the National Library of Sweden. It was trained on slightly less than 3.5 million tokens, with text coming from digitized newspapers, official reports from the Swedish government, legal resources, social media, and Wikipedia in Swedish. They used the same code and hyperparameters as the original BERT (Devlin et al., 2019) model did.

The specific implementation that we are using is the one released on KBLab's HuggingFace repository.[11] Furthermore, we use the BERT for classification class from HuggingFace. It adds a linear layer on top of the base model, with an output for each of the classes. The whole model is then fine-tuned on the training data.

3.4 Evaluation

To measure the biases within the classification task, we use equality of opportunity (Hardt et al., 2016). Equality of opportunity is achieved when the recall between a given class and the rest of the population is equal. This metric is used to minimize false negatives, thus measuring whether any of the groups gets a systemic unfair disadvantage.

In more mathematical terms, if we have the name group A, the recall on its respective diagnostic essays RC_A, and the recall for the rest of the essays on the diagnostic set RC_{-A}, then we can define equality of opportunity for group A as follows:

$$Eq.ofOpp.(A) = RC_A - RC_{-A}$$

A negative value in the metric means that using names from group A in the text of the essay increases the possibility of an unfair disadvantage, while a positive value means that names from that group are less likely to be disadvantaged.

Do note that Hardt et al. (2016) also propose another metric called equalized odds, where we expect both recall and precision to be the same. However, they argue that it is a much stronger requirement and prove that predictors in general cannot be balanced post-hoc to achieve this definition of fairness.

4 Results and Discussion

We can notice from Table 2 that the transformer-based model performs much better than the feature-based model across all evaluation metrics. On top of that, we realized both during training and during inference that BERT was much faster than the feature-based model due to the API calls required to obtain said features.

When looking at the performance on the diagnostic set in Tables 3 and 4, we noticed that changing the names in the text of the essays yielded no change in performance with either of the models. That is, the equality of opportunity of the different groups and subgroups is zero, indicating that the model is not unfair under this metric. Testing with a wider array of names yielded no differences either in terms of class assigned. On a similar note, when checking for biases regarding whether the names were male or female we found no difference in performance.

As mentioned in Section 3.3, we did not expect the feature-based model to show much bias, if at all. This is due to it not using features directly related to the vocabulary.

On the other hand, we expected the transformer-based model to display some sort of bias considering the previous literature on name biases in NLP (see Section 2). This means that ultimately neither the distribution of the demographics in the training set nor the biases in the base BERT model (i.e. intrinsic biases) had any effect on the fairness of the model (i.e. extrinsic bias). A possible direction on which this study could be expanded to would be a thorough analysis of given names present in the vocabulary of the BERT model and seeing whether there is any correlation between how the model behaves for each of these.

These results are consistent with what we would expect from a fair model for AES for second language assessment. That is, we expect it to score the students in terms of their linguistic skills and proficiencies as opposed to other unrelated things.

One of the possible issues that we could have run into were the essays used for the diagnostics dataset. While they represent different CEFR lev-

[10] https://www.kb.se/in-english/research-collaboration/kblab.html

[11] https://huggingface.co/KBLab/bert-base-swedish-cased

Model	Accuracy	F1 Macro	F1 Weighted
Feature-Based	0.25	0.08	0.1
BERT	0.66	0.65	0.65

Table 2: Performance of the models on the test set. Note that the transformer-based architecture fares much better than the feature-base one. Also note that the test set contains unaltered essays, as opposed to the diagnostic set.

	Feature-Based		BERT	
Name Groups	Accuracy	Recall	Accuracy	Recall
Swedish	0.14	0.20	0.86	0.60
Finnish	0.14	0.20	0.86	0.60
Anglo-American	0.14	0.20	0.86	0.60
Arabic	0.14	0.20	0.86	0.60

Table 3: Performance of the models on the diagnostics set. Note that both the accuracy and the recall are the same for all ethnic groups. Also note that the diagnostic set contains the essays with the substituted names, as opposed to the test set.

Name Groups	Feature-Based	BERT
Swedish	0.0	0.0
Finnish	0.0	0.0
Anglo-American	0.0	0.0
Arabic	0.0	0.0

Table 4: Equality of opportunity results for the different name groups chosen. Note that the values are zero for all, meaning that the models do not discriminate based on these names for the essays in the diagnostic set.

els, text genres, and who the name refers to, we still had a small amount of essays to work with. Antoniak and Mimno (2021) note that the choice of seeds for measuring bias can affect the results of such measurements. Thus, using more essays would be good way to verify that our results indeed generalize. However, none of the essays in SweLL-pilot are fit for the criteria we mentioned in Section 3.1 so this would require either gathering new data or generating synthetic data. It is also important to take into account that the size of the diagnostic dataset scales quickly, as it gets 80 new datapoints for each new essay we add.

It is important to note that these results do not mean that neither the base model nor the training data contain biases. They just mean that we did not find biases when using them for the AES task. It has been noted before that intrinsic and extrinsic biases do not necessarily correlate with each other (Goldfarb-Tarrant et al., 2021). That is, just because we did not find biases on our specific task, that does not mean that one can assume that neither Swedish BERT nor the SweLL-pilot are bias-free. That is, we cannot use them for other tasks or applications without worrying about bias or fairness.

5 Conclusions and Future Work

In this work we examined how changing given names within the text of second language learner essays of Swedish affects the CEFR level they are assigned to by the models. We found that changing the names did not change the performance of the model in any noticeable way across four different name groups with twenty names each.

This points to our models learning to differentiate the level of an essay based on linguistic characteristics, as opposed to the kind of personal identifiable information found within the essays, such as given names. Because of this, we think that pseudonymization should be considered as a viable method to allow for research data to be used and shared.

However, it is important to note that these results could vary from language to language and from dataset to dataset. There is no silver bullet to solve the bias issue in NLP, as it is deeply ingrained within human perception and the data we generate, which can lead to unexpected results (Wang et al., 2019-10). Moreover, it is possible that the chosen given names and ehtnic groups could have had an impact on our results, as argued by Antoniak and Mimno (2021). This would be particularly important when considering people coming from regions

under-represented in our data, as they are the most at risk of being the most affected by discrimination, be it from humans or from machines.

There are several directions in which our work could be expanded to. One would be to use more essays for the diagnostic dataset. As mentioned in Section 4, this would require either acquiring new essays or generating synthetic data, both of which can be challenging tasks.

Another possible direction to expand our work to would be to do an in-depth analysis of the given names appearing both in different corpora as well as in the training data of the different models. This would allow us to verify that the lack of perceptible bias we found was not due to the names not appearing on the data.

Both of these could be used as a paving stone to create guidelines on how to generate diagnostic datasets to identify biases in automated essay scoring of second language learner essays. It would be particularly interesting to analyze whether the same patterns hold for different kinds of personal identifiable information, such as other kinds of personal names and places. Moreover, it would be good to check whether this apparent lack of bias is maintained when dealing with several pieces of private information at the same time.

Ethics Statement

Different kinds of data are more likely to contain personal information. This impacts how the data can be used in an ethical way for research. Written consent was obtained during the collection process of the essays from the SweLL-pilot corpus and the data was processed in accordance to the GDPR. The original, non-anonymized data is used strictly within the project, with the real names of the authors of the essays never being disclosed. At the moment in which the data was originally gathered and released, there was no requirement of ethical review.

Special care was put when selecting both the ethnic groups to include and the names belonging to these, as noted in Section 3.2. As mentioned, both of these were chosen to represent some of the most commonly ocurring names in which we would expect AES for second language assessment to occur. While this would showcase any systemic biases that could occur at scale, it ignores under-represented minorities which tend to be the most affected by these kind of things. Thus, it is of utmost importance that if any such system were to be put to use on any potentially life-changing situation, care should be taken to show that even these minorities are assessed in a fair and unbiased manner.

Even though our study strongly points to a lack of biases regarding given names appearing in the text of the essays, any such systems should be continuously monitored to avoid biases appearing seemingly out of nowhere. The use of different datasets and of different methodologies could lead to different results, especially considering how these things might drift over time. Moreover, any high-stakes applications should still have a human-in-the-loop approach so as to ensure that test-takers have access to their rights of explanation and of revision.

Acknowledgements

This work has been possible thanks to the funding of two grants from the Swedish Research Council.

The project *Grandma Karl is 27 years old: Automatic pseudonymization of research data* has funding number 2022-02311 for the years 2023-2029.

The Swedish national research infrastructure *Nationella Språkbanken* is funded jointly by contract number 2017-00626 for the years 2018-2024, as well 10 participating partner institutions.

References

Emilia Aldrin. 2017. Assessing Names? Effects of Name-Based Stereotypes on Teachers' Evaluations of Pupils' Texts. *Names*, 65(1):3–14.

Afra Alishahi, Yonatan Belinkov, Grzegorz Chrupała, Dieuwke Hupkes, Yuval Pinter, and Hassan Sajjad, editors. 2020. *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, Online.

Association of Language Testers in Europe ALTE. 2020. ALTE Principles of Good Practice.

Tracy N Anderson-Clark, Raymond J Green, and Tracy B Henley. 2008. The relationship between first names and teacher expectations for achievement motivation. *Journal of Language and Social Psychology*, 27(1):94–99.

Maria Antoniak and David Mimno. 2021. Bad seeds: Evaluating lexical methods for bias measurement. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages

1889–1904, Online. Association for Computational Linguistics.

American Speech-Language-Hearing Association ASLHA. 2023. Rights and Responsibilities of Test Takers: Guidelines and Expectations.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186. Publisher: American Association for the Advancement of Science Section: Reports.

COE Council of Europe. 2001. *Common European Framework of Reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

EU EU Commission. 2016. *General data protection regulation.* Official Journal of the European Union, 59, 1-88.

David N Figlio. 2003. Names, expectations and black children's achievement. *Unpublished manuscript*.

Gigi Foster. 2008. Names will never hurt me: Racially distinct names and identity in the undergraduate classroom. *Social science research*, 37(3):934–952.

Doreen Gerritzen. 2007. First names of moroccan and turkish immigrants in the netherlands. In Eva Brylla and Mats Wahlberg, editors, *Proceedings of the International Congress of Onomastic Sciences 21, Uppsala August 2002*, pages 120–130. SOFI (Språk- och folkminnesinstitutet, Institute for Dialectology, Onomastics and Folklore Research).

Seraphina Goldfarb-Tarrant, Adam Lopez, Roi Blanco, and Diego Marcheggiani. 2023. Bias beyond English: Counterfactual tests for bias in sentiment analysis in four languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4458–4468, Toronto, Canada. Association for Computational Linguistics.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1926–1940, Online. Association for Computational Linguistics.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Anthony G Greenwald, Mahzarin R Banaji, and Brian A Nosek. 2015. Statistically small effects of the Implicit Association Test can have societally large effects. *American Psychological Association*.

Anthony G Greenwald and Linda Hamilton Krieger. 2006. Implicit bias: Scientific foundations. *California law review*, 94(4):945–967.

Linnea Gustafsson. 2021. Syriska förnamn i sverige. en första kartläggning. In *Navn på minoritetsspråk i muntlige og skriftlige sammenhenger*, volume 99, pages 55–68. Sámi allaskuvla / Sámi University of Applied Sciences, NORNA-förlaget. Conference Name: 49th NORNA-symposium: Minority Names in Oral and Written Contexts in a Multi-Cultural World, Guovdageaidnu (Kautokeino), Norway, 24–25 april, 2019 Publisher: Sámi allaskuvla Accepted: 2022-02-22T08:18:55Z ISSN: 0332-7779 Journal Abbreviation: Minoritehtagielaid namat njálmmálaš ja čálalaš oktavuođain Publication Title: 276.

Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 3323–3331. Curran Associates Inc.

Michael T Kane. 2001. Current concerns in validity theory. *Journal of educational Measurement*, 38(4):319–342.

Anastassia Loukina, Nitin Madnani, and Klaus Zechner. 2019. The many dimensions of algorithmic fairness in educational applications. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–10.

Nitin Madnani, Anastassia Loukina, Alina Von Davier, Jill Burstein, and Aoife Cahill. 2017. Building better open-source tools to support fairness in automated scoring. In *Proceedings of the first ACL workshop on ethics in natural language processing*, pages 41–52.

Ylva Malm and Pontus Zetterström. 2007. *Kevins konnotationer - skillnader i högstadielärares associationer till tio olika förnamn*. Örebro University.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with words at the national library of sweden – making a swedish bert.

Marta Marchiori Manerba, Riccardo Guidotti, Lucia Passaro, and Salvatore Ruggieri. 2022. Bias discovery within human raters: A case study of the jigsaw dataset. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 26–31, Marseille, France. European Language Resources Association.

Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao Huang, and Shomir Wilson. 2023. Nationality bias in text generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 116–122, Dubrovnik, Croatia. Association for Computational Linguistics.

Ildikó Pilán, Elena Volodina, and Torsten Zesch. 2016. Predicting proficiency levels in learner writings by transferring a linguistic complexity model from expert-written coursebooks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2101–2111, Osaka, Japan. The COLING 2016 Organizing Committee.

Carsten Roever and Tim McNamara. 2006. Language testing: The social dimension. *International Journal of Applied Linguistics*, 16(2):242–258.

Sandra Sandoval, Jieyu Zhao, Marine Carpuat, and Hal Daumé III. 2023. A rose by any other name would not smell as sweet: Social bias in names mistranslation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3933–3945, Singapore. Association for Computational Linguistics.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Harini Suresh and John Guttag. 2021. A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21, New York, NY, USA. Association for Computing Machinery.

Elena Volodina. 2024. On two SweLL learner corpora – SweLL-pilot and SweLL-gold. *Huminfra Conference*, pages 83–94.

Elena Volodina, Ildikó Pilán, Ingegerd Enström, Lorena Llozhi, Peter Lundkvist, Gunlög Sundberg, and Monica Sandell. 2016a. SweLL on the rise: Swedish learner language corpus for European reference level studies. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 206–212, Portorož, Slovenia. European Language Resources Association (ELRA).

Elena Volodina, Ildikó Pilán, and David Alfter. 2016b. Classification of Swedish learner essays by CEFR levels. In *CALL communities and culture – short papers from EUROCALL 2016*, pages 456–461. Research-publishing.net.

Jun Wang, Benjamin Rubinstein, and Trevor Cohn. 2022. Measuring and mitigating name biases in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2576–2590, Dublin, Ireland. Association for Computational Linguistics.

Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. 2019-10. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5309–5318. ISSN: 2380-7504.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

A Essays Used for Diagnostics Purposes

The following are the IDs for the essays chosen for diagnostic purposes:

- S143ST18
- S147ST18
- S42ST9
- S53ST12
- W13WT2
- W53WT5
- W2WT2

B Lists of Names Used

This appendix contains Tables 5, 6, 7, and 8. These four tables show the names used for each group in this study.

Female	Male
Anna	Lars
Eva	Mikael
Maria	Anders
Karin	Johan
Sara	Erik
Christina	Karl
Lena	Per
Emma	Olof
Kerstin	Nils
Marie	Jan

Table 5: List with the Swedish names chosen for this study, as specified in Section 3.2.

Female	Male
Hannele	Juhani
Marjatta	Eino
Maarit	Olavi
Annikki	Antero
Aurora	Tapani
Aino	Kalevi
Helmi	Tapio
Ilona	Matti
Minna	Ilmari
Sari	Onni

Table 6: List with the Finnish names chosen for this study, as specified in Section 3.2.

Female	Male
Mary	Kevin
Patricia	James
Jennifer	Charles
Nancy	John
Betty	Matthew
Barbara	Anthony
Susan	William
Jessica	Donald
Ashley	Steven
Karen	Brian

Table 7: List with the Anglo-American names chosen for this study, as specified in Section 3.2.

Female	Male
Fatima	Muhammad
Hala	Ali
Amal	Ahmed
Mariam	Ibrahim
Hiba	Hassan
Huda	Mahmoud
Khadija	Omar
Mirna	Abdullah
Samira	Ismail
Fatemeh	Hamza

Table 8: List with the Arabic names chosen for this study, as specified in Section 3.2.