Toward Consistency in Writing Proficiency Assessment: Mitigating Classification Variability in Developmental Education

Miguel Da Corte[1,2][a] and Jorge Baptista[1,2][b]
[1]*University of Algarve, Faro, Portugal*
[2]*INESC-ID Lisboa, Lisbon, Portugal*

Keywords: Developmental Education (DevEd), Automatic Writing Assessment Systems, English (L1) Writing Proficiency Assessment, Natural Language Processing (NLP), Machine-Learning (ML) Models.

Abstract: This study investigates the adequacy of Machine Learning (ML)-based systems, specifically ACCUPLACER, compared to human rater classifications within U.S. Developmental Education. A corpus of 100 essays was assessed by human raters using 6 linguistic descriptors, with each essay receiving a skill-level classification. These classifications were compared to those automatically generated by ACCUPLACER. Disagreements among raters were analyzed and resolved, producing a gold standard used as a benchmark for modeling ACCUPLACER'S classification task. A comparison of skill levels assigned by ACCUPLACER and humans revealed a "weak" Pearson correlation ($\rho = 0.22$), indicating a significant misplacement rate and raising important pedagogical and institutional concerns. Several ML algorithms were tested to replicate ACCUPLACER'S classification approach. Using the Chi-square (χ^2) method to rank the most predictive linguistic descriptors, Naïve Bayes achieved 81.1% accuracy with the top-four ranked features. These findings emphasize the importance of refining descriptors and incorporating human input into the training of automated ML systems. Additionally, the gold standard developed for the 6 linguistic descriptors and overall skill levels can be used to (i) assess and classify students' English (L1) writing proficiency more holistically and equitably; (ii) support future ML modeling tasks; and (iii) enhance both student outcomes and higher education efficiency.

1 INTRODUCTION AND OBJECTIVES

This study examines the adequacy of a machine-learning-based placement system, ACCUPLACER, and compares it to the classifications made by human raters within the context of U.S. Developmental Education (DevEd). By focusing on placement accuracy, this paper evaluates how reliable this system is and how effectively it supports higher education institutions in placing students into appropriate courses.

DevEd courses are designed to support students who are not yet college-ready, particularly by enhancing their English writing skills, ensuring they can gain access to and successfully participate in academic programs. Colleges typically assign students to either DevEd or college-level courses based on how they perform on standardized entrance exams. These decisions have significant consequences, as students with lower scores may need to complete one or two semesters of developmental, sometimes referred to as 'remedial,' coursework (Bickerstaff et al., 2022). Furthermore, the major concern among higher education institutions is the fact that current placement assessments are often unreliable in predicting students' performance, leading to incorrect placement for up to one-third of those who take the tests (Ganga and Mazzariello, 2019).

According to the National Center for Education Statistics (NCES)[1] approximately 3.6 million students graduated U.S. high school during the 2021-2022 academic year, with approximately 62%, ages 16 to 24, enrolling in colleges or universities as recently reported by the United States Bureau of Labor Statistics[2]. At Tulsa Community College[3], where this study was carried out, around 30% of incoming students are deemed not college-ready in more than

[a] https://orcid.org/0000-0001-8782-8377
[b] https://orcid.org/0000-0003-4603-4364

[1] https://nces.ed.gov
[2] https://www.bls.gov
[3] https://www.tulsacc.edu

Da Corte, M. and Baptista, J.
Toward Consistency in Writing Proficiency Assessment: Mitigating Classification Variability in Developmental Education.
DOI: 10.5220/0013353900003932
In *Proceedings of the 17th International Conference on Computer Supported Education (CSEDU 2025) - Volume 2*, pages 139-150
ISBN: 978-989-758-746-7; ISSN: 2184-5026

one subject - English (reading and writing) and Math, highlighting the need for reliable placement mechanisms. Inaccurate placement in DevEd means students are assigned to courses beyond their writing abilities or to those that underestimate their skills, both of which have significant consequences for student outcomes and institutional effectiveness (Hughes and Li, 2019; Link and Koltovskaia, 2023; Edgecombe and Weiss, 2024).

This study aims to address these challenges by analyzing a corpus of 100 essays from college-intending students, evaluated according to six linguistic descriptors used by ACCUPLACER. The essays were classified by six human raters into one of three skill levels (DevEd Level 1, Level 2, or College Level) based on guidelines developed by two linguists using these descriptors. Subsequently, raters assessed the global level of the essays. This process unfolded in two phases over a period of one month.

A critical issue underlying this classification process lies in the nature of the descriptors themselves. Defined as high-order constructs, these descriptors encapsulate broad linguistic competencies but lack the granularity necessary to capture finer distinctions in writing proficiency. This study interrogates the extent to which these descriptors, particularly those associated with higher-order thinking, may require refinement to enhance assessment consistency and reliability. By examining how these constructs influence essay classification, the study seeks to provide insights into the alignment between linguistic features and skill-level assignments, ultimately informing the refinement of automated and human-scored placement decisions.

Given the background and context provided, this study has four objectives: (i) identify specific areas of disagreement in the application of six language-proficiency descriptors by human raters; (ii) compare the skill levels assigned by human raters with those automatically produced by ACCUPLACER; (iii) analyze and address discrepancies between proficiency descriptors and skill-level assignments; and (iv) develop a gold-standard corpus with a focus on achieving high inter-rater agreement.

Regarding its applicability to other languages, this study specifically focuses on English-language placement within U.S. community colleges. While extending the methodology to other languages is valuable, it falls outside this study's focus.

2 RELATED WORK

Current guidelines for DevEd assessment and placement vary widely across institutions and states, leading to inconsistencies in defining and categorizing proficient writing (Kopko et al., 2022). Automated systems like ACCUPLACER aim to reduce human biases and improve consistency in the assessment process (Link and Koltovskaia, 2023), but often lack the detailed insights—such as learners' early writing patterns—needed to shape effective instructional practices for developing writers (Da Corte and Baptista, 2024a).

Research calls for higher education institutions to turn to more comprehensive assessment tools to evaluate writing competencies, particularly for college readiness (Gallardo, 2021). As noted by (Lattek et al., 2024), "not every key competency can be assessed and measured using the same assessment method or instrument [...]; however, a suitable assignment or systematization is currently lacking," highlighting the need for refined scoring systems to ensure equitable and efficient access to higher education through linguistic development opportunities. Leveraging natural language processing (NLP) techniques (Link and Koltovskaia, 2023) could address these gaps and mitigate concerns about the accuracy and validity of current assessment methods (Barnett et al., 2020; Perelman, 2020).

Previous studies have aimed to detect textual patterns through computational methods, e.g., 'narrative style,' 'simple syntactic structures,' 'cohesive integration' (Dowell and Kovanovic, 2022), and to identify the linguistic features that are most predictive of accurate placement in Developmental Education (DevEd) (Da Corte and Baptista, 2024a). By focusing on descriptive features that better reflect native English proficiency, automatic classification systems have shown improvements in placement accuracy (Da Corte and Baptista, 2022), offering valuable insights into how to prepare students more effectively both linguistically and academically (Bickerstaff et al., 2022; Giordano et al., 2024).

Building on these findings, (Sghir et al., 2023; Duch et al., 2024) examined the use of ML algorithms to enhance the assessment of students' written productions and predict their performance and placement based on observable writing patterns (dos Santos and Junior, 2024). The overall findings indicate that well-known ML algorithms, such as Naïve Bayes, Neural Networks, and Random Forest, among others, are highly effective at pattern detection and feature selection (Hirokawa, 2018), and at predicting

students' outcomes, with classification accuracy rates above 90% (Duch et al., 2024).

As a result, refining proficiency descriptors, addressing discrepancies in their application by human raters and incorporating more descriptive linguistic features into the training of systems like ACCUPLACER could improve skill-level classification and better support DevEd students and faculty through learner corpora (Götz and Granger, 2024). Furthermore, developing a reliable, gold-standard corpus with high inter-rater agreement will establish a foundation for future assessments, supporting the overarching goals of this study.

3 METHODOLOGY

Through a systematic sampling method in accordance with the Institution's Review Board (IRB) protocols (ID #22-05[4]), this study ensured ethical and fair participant selection. It specifically focused on individuals who are educationally disadvantaged and adhered to guidelines that address the unique academic challenges faced by this group.

A rigorously selected corpus of 100 essays was assessed using six specific linguistic descriptors. Two linguists developed and tested classification guidelines based on these descriptors, and the essays were rated by trained human raters using a simplified 4-point Likert scale. Additionally, a pilot was conducted to validate the annotation approach, resulting in a comprehensive human-rated dataset benchmark. This golden standard dataset was then compared against the automated classification produced by ACCUPLACER.

3.1 Corpus

The current study builds on a previous classification task (Da Corte and Baptista, 2024b) that utilized a carefully selected corpus of 100 essays (27,916 tokens total)[5], written by college-intending students during the 2021-2023 academic years. Extracted from a larger pool of 290 essays within the standardized entrance exam database, these texts provided a robust foundation for assessing writing proficiency. Written in the Institution's proctored testing center without access to the Internet or editing tools, the essays were based on diverse prompts, such as the value of history, the acquisition of money, and the results of deception, designed to elicit analytical and reflective responses.

The selection process ensured that the corpus was representative of students' DevEd placement levels, categorized by ACCUPLACER as Level 1 or 2. The essays were balanced in level, as the classification by level was the target metric, and averaged 260 tokens each. Although small, this corpus serves as a critical foundation for analyzing the language proficiency of community college students. Demographic information (e.g., gender, race) was ignored at this stage.

This study's corpus specifically targets non-college level classifications to evaluate readiness for college and identify students needing DevEd support. Given the variability in DevEd guidelines across U.S. institutions, focusing on this segment is essential to refine placement accuracy through automated systems like ACCUPLACER, ensuring equitable access for academically underprepared students.

3.2 Classification Task Set-up

Before the classification task began, two linguists developed a set of guidelines[6], drawing on the six textual descriptors used in the ACCUPLACER assessment (The College Board, 2022). These guidelines were then tested in a pilot study, where a small selection of texts from the same exam database and timeframe was annotated. The focus was on identifying and categorizing relevant linguistic descriptors within DevEd.

In this classification task, all text samples (100) were randomly assigned to six pairs of raters, ensuring that each essay was reviewed by at least two individuals. To manage the substantial workload, the task was divided into two batches of 50 essays each, completed over the course of one month. Prior to this, the raters, who were familiar with DevEd and ACCUPLACER course placement in higher education institutions, received comprehensive training led by one of the authors of the annotation guidelines. This training covered the task's expectations, ethical considerations, an explanation of the guidelines, and the overall annotation procedures.

Following the training, the raters proceeded with the assessment, which involved two key steps:

(i) assessing each essay according to six specific textual criteria, defined in the ACCUPLACER manual (The College Board, 2022, p. 27):

(1) *Mechanical Conventions* (MC);
(2) *Sentence Variety and Style* (SVS);
(3) *Sentence Development and Support* (DS);
(4) *Organization and Structure* (OS);
(5) *Purpose and Focus* (PF); and
(6) *Critical Thinking* (CT).

[4]https://www.tulsacc.edu/irb

[5]Corpus dataset to be made available after paper publication.

[6]https://gitlab.hlt.inesc-id.pt/u000803/deved/

For this evaluation, a simplified 4-point Likert scale[7] was applied to each criterion:

0 for *deficient*;
1 for *below average*;
2 for *above average*; and
3 for *outstanding*.

(ii) assigning an overall classification to each essay. Essays were classified based on definitions specifically adapted from the Institution's official course descriptions and curriculum:

DevEd Level 1 if essays demonstrated a need for improvement in general English usage, including grammar, spelling, punctuation, and the structure of sentences and paragraphs;

DevEd Level 2 if essays required targeted support in specific aspects of English, such as sentence structure, punctuation, editing, and revising; or

College level if essays were written accurately and displayed appropriate English usage at the college academic level.

4 EXPLORING VARIATIONS IN HUMAN RATER EVALUATIONS

4.1 Linguistic Descriptors

Each of the 100 essays received two scores, one from each rater, for the 6 linguistic descriptors analyzed. The assessments were completed on schedule, with raters self-reporting an average of 11 minutes per writing sample. Table 1 presents the number of essays that received differing scores in one or more descriptors.

Table 1: Essays that received differing scores in one or more descriptors.

Differing scored descriptors	Number of Essays
Essays with 0 differing scored descriptors	37
Essays with 1 differing scored descriptor	26
Essays with 2 differing scored descriptors	17
Essays with 3 differing scored descriptors	11
Essays with 4 differing scored descriptors	9
Essays with 5 or 6 differing scored descriptors	0
Total	100

Approximately 1/3 of the essays (37) received equal scores across all descriptors, indicating strong agreement between raters and suggesting a consistent application of the guidelines in these cases. In contrast, approximately 2/3 of the essays (63) had differing scores in one or more descriptors, hinting at some interpretative issues with the definitions provided.

Within a significant portion of this subset (63), 26 text samples had discrepancies in only 1 descriptor, while 17 of them exhibited differences in 2 descriptors. These essays likely represent cases where raters had minor disagreements, possibly due to subjective interpretation of specific descriptors. A total of 20 essays (combining those with 3 and 4 differing descriptors) revealed more pronounced disagreements among raters, indicating that scoring consistency decreases when assessing multiple linguistic aspects. No essays had discrepancies in 5 or 6 descriptors.

Upon close inspection of these 63 texts, three key themes are noted. Having a corpus of 100 essays and 6 linguistic descriptors results in a dataset with a total of 600 possible data points where differences between raters may occur. The focus here is on cases where Raters 1 and 2 provided contradictory scores, crossing the positive/negative boundary—where values of 0 or 1 signal a deficient or below-average text, and values of 2 or 3 signal an above-average or outstanding text.

Out of these 600 data points: 105 cases involved a one-point difference (e.g., 1 to 2) on the 4-point Likert scale; 23 cases involved a two-point difference (e.g., 0 to 2, 1 to 3); and 1 case, concerning the *Development and Support* descriptor, involved a three-point difference (3 to 0). In the remaining 471 instances, although some differing scores were observed within the same descriptors, they fell within either the negative (0, 1) or positive (2, 3) boundaries without crossing the +/- threshold.

To obtain a final score for each of the respective linguistic descriptors for the 63 essays included in Table 1, a third independent rater (Rater 3) was consulted. Rater 3 had not previously seen or evaluated the essays and provided an additional perspective to the assessment process. Results are included in Appendix 7.

4.1.1 Understanding Descriptor Discrepancies: Toward a Gold Standard

Although the descriptors used by ACCUPLACER align with general academic writing standards, they are not specifically tailored to the unique needs of DevEd students. These high-level descriptors are challenging to apply consistently, even for trained human raters, and based on current DevEd literature, lack grounding in detailed linguistic research addressing the specific requirements of DevEd contexts.

[7]This scale simplifies the more complex 8-point Likert scale currently used by ACCUPLACER (The College Board, 2022, p. 24).

The themes explored in Section 4.1 suggest that some descriptors, particularly those related to higher-order thinking, may be overly broad and could benefit from refinement. Narrowing their scope or providing more detailed guidelines could reduce discrepancies and improve assessment consistency. Table 2 focuses on analyzing which linguistic descriptors showed the greatest divergence in raters' scores, providing a basis for targeted improvements.

Table 2: Essays with differing scores per descriptor.

Descriptor	Number of Essays
Mechanical Conventions	12
Organization and Structure	16
Critical Thinking	16
Purpose and Focus	19
Sentence Variety and Style	29
Development and Support	37

During the pilot study mentioned in Section 3, refinements were made to the descriptors *Mechanical Conventions* and *Organization and Structure* to enhance their clarity and applicability. These efforts resulted in *Mechanical Conventions* achieving the fewest score discrepancies (12) among the raters. This improvement is attributed to specific enhancements made by two linguists to the descriptor's definition in the ACCUPLACER manual, including the introduction of a numerical scale for evaluating misspelled words:

0 (deficient) for 15 or more misspelled words;
1 (below average) for 8-14 misspelled words;
2 (above average) for 1-7 misspelled words; and
3 (outstanding) for no misspelled words.

However, unlike the refined scale for spelling errors, other grammatical features that could be incorporated into the *Mechanical Conventions* descriptor—such as word omission, word repetition, subject-verb disagreement (e.g., *we was* instead of *we were*), and punctuation misuse—lack similar metrics. These features will be considered in future refinements, as the ACCUPLACER manual provides no guidance on their inclusion in the assessment process.

Next are *Organization and Structure* and *Critical Thinking*, both with 16 essays each receiving differing scores. Minor enhancements were made to the *Organization and Structure* descriptor to include the fundamentals of paragraph composition: a clear introduction with a thesis statement, supporting statements, and a conclusion. This adjustment is believed to have enhanced the objectivity of the assessment. Texts relying on single lines or simple, non-multilayered paragraphs were rated as deficient, whereas those featuring well-structured paragraphs—with a clear introduction, at least two supporting statements, and a conclusion—were considered outstanding.

No enhancements were made to the *Critical Thinking* descriptor nor the remaining descriptors, as the intent was to closely assess the ACCUPLACER's classification criteria, and its reproducibility using human raters. *Critical Thinking*, despite being a complex and abstract feature, was measured based on constructs like *fairness*, *relevance*, *precision*, and *logic*, among others. Interestingly, it had the same number of essays scored differently as one of the descriptors whose definition was enhanced - *Organization and Structure*.

Under *Purpose and Focus*, 19 essays received differing scores. This feature evaluates how effectively a text presents information in a unified, coherent, and consistent manner. It also addresses the concept of *relevance*, thus partially overlapping with *Critical Thinking*. These aspects often require robust NLP tools for accurate detection and objective assessment.

The descriptors with the most scoring discrepancies were *Sentence Variety and Style* (29) and *Development and Support* (37). For *Sentence Variety and Style*, constructs such as *sentence length*, *vocabulary variety*, and *voice* are considered and can be easily quantified. For instance, sentence length exceeding 15 to 17 words—typically considered the standard for improved readability and comprehension (Matthews and Folivi, 2023)[8]—could serve as a basis for distinguishing between deficient and below-average texts and developing a numerical scale for this particular descriptor.

Regarding *Development and Support*, factors such as *point of view*, *coherent arguments*, and *evidence* are evaluated. While certain aspects, like the frequency of keywords introducing examples (e.g., *as proof*, *to give an idea*, *for example*) or reasoning (e.g., *because*, *although*, *consequently*), can be objectively measured, assessing a writer's point of view presents a significant challenge. This difficulty arises early in some students' academic journey, as articulating a coherent viewpoint often requires a level of maturity and experience gained through their studies.

While refining linguistic descriptors is essential for improving consistency and accuracy in rater assessments, it is equally important to consider how these descriptors contribute to the overall skill level classification. By examining how descriptors are applied to determine skill levels, discrepancies between human raters and ACCUPLACER classifications can be better understood, and that is the purpose of Section 4.2.

[8]https://readabilityguidelines.co.uk/

4.2 Skill Level

The overall skill level classification of the 100 essays is as follows: 68 essays received identical classification levels from both Rater 1 and Rater 2, while 32 essays were classified differently. For these 32 essays where discrepancies existed between human classifications, Rater 3, already mentioned in Section 4.1, was also tasked with independently providing a supplementary assessment to resolve the differences and confirm the skill level of the text samples. Still, for 2 essays, no agreement was reached, as each rater assigned a different skill level (Level 1, 2, and College Level). The final score was therefore determined by averaging the three ratings.

A custom scale based on the minimum and maximum average scores across all 6 linguistic descriptors was developed to further examine these differences. This scale provided a systematic framework for classifying texts into DevEd Level 1, DevEd Level 2, and College Level proficiency, with the following ranges:

Level 1 = [0.000 - 1.499];
Level 2 = [1.500 - 2.499]; and
College Level = [2.500 - 3.000].

Using this scale, the discrepancies noted in the assigned DevEd levels are summarized in Table 3.

Table 3: Discrepancies among assigned DevEd Levels.

Scale	Levels	Level 1	Level 2	College Level	Total
0.000 - 1.499	Level 1	61	**7**	0	68
1.500 - 2.499	Level 2	**11**	20	0	31
2.500 - 3	College Level	0	**1**	0	1
Total		72	28	0	100

From this summary, it was observed that: 7 texts with below-average and deficient scores (across the 6 linguistic descriptors) were assigned to DevEd Level 2; 11 texts with above-average and outstanding scores had been placed in DevEd Level 1; and 1 text with above-average and outstanding scores could have been placed in College-level writing but was deemed as a Level 2 text. The final human-assigned skill-level classifications are: 72 texts at Level 1 and 28 texts at Level 2.

Following this exploration of variability among human raters and the resulting assessment of all 100 text samples, a gold standard was established for non-college-level DevEd writing for the 6 linguistic descriptors and skill levels. Uniquely curated through human rater input, this gold standard is detailed in Appendix 7 and represents a significant advancement in creating a carefully vetted and reliable dataset for classification. It establishes a publicly available standard for this population that, to the best of the authors' knowledge, has not previously existed. Furthermore, this gold standard is essential for modeling the classification process using ML techniques and for understanding the role of each descriptor in refining the DevEd placement process, as detailed in Section 6.

5 INTER-RATER RELIABILITY AND QUALITY ASSURANCE

Based on the assessment results for the 6 linguistic descriptors and skill level classification, a rigorous quality assurance protocol was followed to evaluate inter-rater reliability, ensuring consistency and accuracy throughout the classification process.

Krippendorff's Alpha (K-alpha) inter-rater reliability coefficients were computed using the ReCal-OIR[9] tool (Freelon, 2013) for ordinal data. The aim was to analyze the level of agreement among the pair of raters for all 6 descriptors and the skill level. For the interpretation of K-alpha scores, the following agreement thresholds and interpretation guidelines, set forth by (Fleiss and Cohen, 1973), were followed:

below 0.20 - *slight* (Sl);
between 0.21 and 0.39 - *fair* (F);
between 0.40 and 0.59 - *moderate* (M);
between 0.60 and 0.79 - *substantial* (Sb);
above 0.80 - almost perfect (P);

Table 4 presents the results using both the granular 0-3 Likert scale, already presented in Section 3.2, and a binary scale, where 0 represents deficient or below-average scores, and 1 represents above-average or outstanding scores. The K-alpha interpretation thresholds are also provided. Items in bold indicate the top 3 linguistic descriptors with the highest reliability scores for each scale.

Table 4: K-Alpha scores for raters across 6 linguistic descriptors and skill levels: Likert vs. Binary scales.

Descriptors	0 - 3 scale	Interp.	0 - 1 scale	Interp.
Mechanical Conventions	**0.566**	M	**0.522**	M
Sentence Variety and Style	**0.396**	F	**0.352**	F
Development and Support	0.303	F	0.232	F
Organization and Structure	**0.433**	M	0.276	F
Purpose and Focus	0.388	F	0.213	F
Critical Thinking	0.361	F	**0.379**	F
Skill Level	0.425	M	0.413	M

K-alpha scores were anticipated to fall within the *slight* to *fair* agreement range, given the high-order nature and complexity of the linguistic descriptors as designed by ACCUPLACER. Results confirmed *fair* agreement for the descriptors *Sentence Variety and Style* (0.396), *Development and Support* (0.303), *Purpose and Focus* (0.388), and *Critical Thinking* (0.361)

[9]https://dfreelon.org/utils/recalfront/recal-oir/

on the 0–3 Likert scale. In contrast, *Mechanical Conventions* (0.566) and *Organization and Structure* (0.433) achieved *moderate* agreement.

On the binary scale, *Mechanical Conventions* was the only descriptor to achieve *moderate* agreement, representing the highest K-alpha scores overall across both scales. Skill level classification also demonstrated *moderate* agreement (0.425 on the Likert scale; 0.413 on the binary scale). The comparison between the binary and Likert scales revealed a "strong positive" Pearson correlation coefficient (Cohen, 1988) of $\rho = 0.754$.

The moderate scores for *Mechanical Conventions*, achieved, in part, due to the numerical scale for typos introduced, along with enhanced guidelines for *Organization and Structure* and self-developed skill level definitions aligned with the DevEd curriculum, suggest that further refining linguistic descriptors with specific, measurable criteria could lead to more objective and consistent evaluations, ultimately improving the placement process for DevEd students.

5.1 An Additional Quality Assurance Measure

Spearman rank correlation coefficients (ρ_s) were calculated based on the two raters' scores for each linguistic descriptor to assess their relative ranking across essays. Results are summarized in Table 5 and interpreted as follows (Mukaka, 2012):

0.90 to 1.00 *Very high positive correlation* (VHP);
0.70 to 0.90 *High positive correlation* (HP);
0.50 to 0.70 *Moderate positive correlation* (M);
0.30 to 0.50 *Low positive correlation* (LP);
0.00 to 0.30 *Negligible correlation* (N).

Table 5: Spearman rank correlation coefficients (ρ) for all 6 linguistic descriptors.

Rank	Linguistic Descriptor	Spearman Correlation	Interp.
1	Mechanical Conventions	0.632	M
2	Organization and Structure	0.521	M
3	Purpose and Focus	0.489	LP
4	Sentence Variety and Style	0.486	LP
5	Critical Thinking	0.471	LP
6	Development and Support	0.406	LP

Mechanical Conventions and *Organization and Structure* were the two descriptors with moderate positive correlation, suggesting a greater level of agreement between the two raters when evaluating these descriptors. The other four descriptors had similar low positive (ρ) scores, which can be attributed to their complexity and broadness, such as assessing the depth of ideas, the appropriateness of sentence structure, or the persuasiveness of an argument. These results seem to confirm the need for further refinements of these descriptors to improve inter-rater reliability.

6 MODELING ACCUPLACER'S CLASSIFICATION

This section investigates how the classification task performed by ACCUPLACER compares to that of human annotators when they apply the same purported linguistic criteria used as annotation guidelines. In alignment with the objectives of this study, outlined in Section 1, the experiments discussed here aim to evaluate the effectiveness of ACCUPLACER's criteria and assess whether human raters can consistently apply them. Eventually, the ultimate goal is to support a more systematic placement of students by enhancing such an automatic classification system.

It is important to note that ACCUPLACER's exact classification process is not publicly detailed, which presents a significant limitation. The system's classification operates as a "black box," and while a summary report is generated after the writing task assessment is completed, it does not detail scoring decisions, hindering educators and students from understanding why certain texts are classified as below college-level. This restricts opportunities for meaningful feedback and targeted improvements. To address these challenges and evaluate ACCUPLACER's performance, the gold standard scores established in Section 4 by human raters were used as a benchmark for modeling the system's DevEd classification task through Machine Learning (ML) experiments.

While ACCUPLACER's skill level classification used for the corpus sampling was balanced (50 essays per DevEd level), the human raters' classification results showed a different scenario (72 essays in DevEd Level 1 and 28 in DevEd Level 2). To further illustrate the discrepancies between ACCUPLACER and human classifications, a confusion matrix (Table 6) provides a detailed breakdown of specific instances where ACCUPLACER encounters challenges in accurately predicting proficiency levels.

Table 6: Accuplacer vs. Human Assessment: 100 Texts.

	Accuplacer L1	Accuplacer L2	Sum
Human L1	50	22	72
Human L2	0	28	28
Sum	50	50	100

Results from this matrix suggest a 22% misplacement rate by ACCUPLACER, where students were placed in DevEd Level 2 despite still needing significant development in their writing skills. The 22% error rate indicates that approximately 1 in 5 stu-

dents could be placed at an inappropriate level. This misclassification has serious pedagogical and institutional implications as (i) students placed below their skill level may face disengagement or frustration from redundant material, and (ii) students placed above their level might struggle, leading to lower success rates and higher attrition.

ACCUPLACER'S limitations and misalignments with human classifications were further validated through Pearson coefficient calculations, yielding a "weak" correlation of $\rho = 0.222$ and a comparable "slight" Krippendorff's Alpha inter-rater reliability coefficient of $k = 0.164$.

Before proceeding with the ML experiments outlined in Section 6.1, random resampling was conducted for Level 1 texts to ensure that the dataset was balanced according to the skill levels attributed by human raters. A total of 56 text samples (28 for each level—Levels 1 and 2) were then used.

6.1 Improving Placement Accuracy Through Machine Learning

The data mining tool ORANGE (Demšar et al., 2013)[10] was selected for analysis and modeling due to its comprehensive suite of popular and commonly used machine learning algorithms. The ORANGE toolkit website provides detailed documentation on classifier selection, definitions, and configurations, facilitating the evaluation of ACCUPLACER's performance relative to human classifications without the need to develop new models.

The workflow as shown in Figure 1 can be described as follows: the data is imported into Orange using the CSV File Import widget - one line per essay, 6 columns for the descriptors, plus a column with the overall skill level, used as the target variable. Data is then passed into the DATA SAMPLER widget, which, considering the small size of the sample, was configured to partition it for 3-fold cross-validation, leaving 2/3 (37 instances) for training and 1/3 (19 instances) for testing purposes. Due to the dataset's configuration, stratified cross-validation is not feasible. The TEST & SCORE widget was then used to determine the best-performing model.

Four, commonly used learning algorithms were chosen for their established performance in similar text classification tasks: Decision Tree (DT), Random Forest (RF), Naïve Bayes (NB), and Neural Network (NN). These algorithms were tested using the default hyperparameters provided by ORANGE. Classification Accuracy (CA) was the primary metric for analysis, while Area Under the Curve (AUC) was employed to differentiate between models with *ex-aequo* CA values.

While previous experiments conducted with the same corpus included six additional algorithms available in ORANGE—Adaptive Boosting, CN2 Rule Induction, Gradient Boosting, k-Nearest Neighbors, Logistic Regression, and Support Vector Machine—this study focused on the four models that delivered the most significant results in this context.

The results from this first part of the experiment are shown in Table 7.

Table 7: ML Algorithm CA w/ 6 linguistics descriptors.

Model	AUC	CA
Random Forest	**0.850**	**0.757**
Neural Network	0.741	**0.757**
Tree	0.722	0.730
Naïve Bayes	0.788	0.703

The order of the algorithms based on their CA scores is as follows: RF > NN > DT > NB. While RF and NN had *ex-aequo* CA scores, RF showed a higher AUC. The RF algorithm is frequently used in similar text classification tasks (Huang, 2023), thus this result is within expectations. The performance of the remainder algorithms is also not very far behind. A CA score of 0.757 suggests that, when using the 6 linguistic descriptors and the corresponding human-annotated scores replicating ACCUPLACER'S classification task, about 3 out of 10 students are inaccurately placed in DevEd courses. This CA value is comparable to the 0.727 CA baseline achieved by RF in a previous DevEd experiment (Da Corte and Baptista, 2024c). In that study, a large set of linguistic features were automatically extracted using Natural Language Processing (NLP) tools like CTAP[11] (Chen and Meurers, 2016), with approximately 300 focused on lexical patterns, such as lexical density and richness, and syntactic patterns, including syntactic complexity and referential cohesion, among others. Although this feature set is broader, the 0.727 CA baseline achieved with RF is only slightly lower than the 0.757 CA noted in this paper using just the six (slightly enhanced) descriptors from ACCUPLACER. This result highlights that, while ACCUPLACER'S constructs can be abstract and difficult for humans to apply in text assessment, they present an opportunity for increased accuracy when descriptors are enhanced with human input. The need for further refinement and improvement remains clear when compared to the more feature-rich approach in the earlier study.

[10] https://orangedatamining.com/

[11] http://sifnos.sfs.uni-tuebingen.de/ctap/

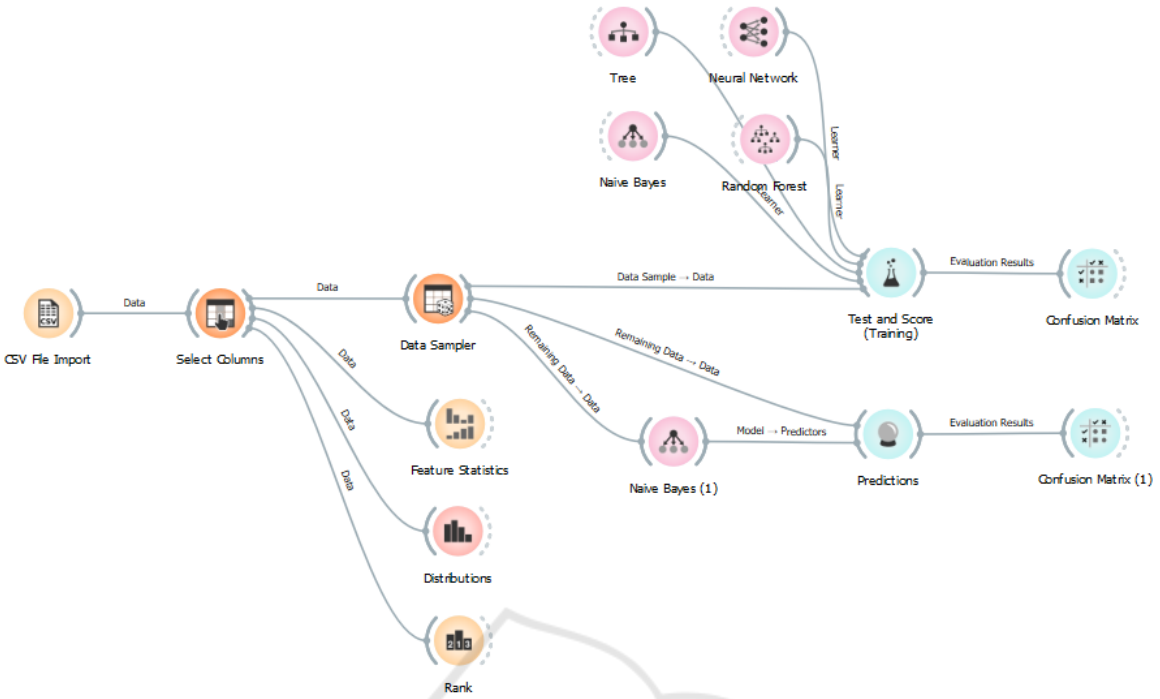

Figure 1: Orange Workflow Configuration for Model Training and Testing.

To examine how the ACCUPLACER'S descriptors contribute to the classification process, the built-in Information Gain and Chi-square (χ^2) ranking methods of Orange's Rank widget were compared to score those features. Figure 2 provides a visual representation of how the features ranked with both methods.

	#	Info. gain	χ^2
1	Development & Support	0.273	6.870
2	Organization & Structure	0.338	6.870
3	Mechanical Conventions	0.159	5.062
4	Sentence Variety & Style	0.218	4.263
5	Critical Thinking	0.187	4.050
6	Purpose & Focus	0.182	3.767

Figure 2: Descriptors ranked by Information Gain and Chi-square (χ^2) methods.

Development and Support and *Organization and Structure* ranked the highest with both ranking methods. As for *Mechanical Conventions* and *Sentence Variety and Style* are also the next highest ranking but in reverse order. *Critical Thinking* and *Purpose and Focus* ranked the lowest with both methods. The "very strong" Pearson correlation coefficient between the two scoring methods ($\rho = 0.824$) led to the adoption of the Chi-square (χ^2).

Using the ranking for feature selection, the top 4 descriptors, *Development and Support*, *Organization and Structure*, *Mechanical Conventions*, and *Sentence Variety and Style* were then used to classify the text samples again. This should prevent overfitting and enable the models to generalize better to unseen data. The results of this second experiment are summarized in Table 8.

Table 8: ML Algorithm CA with Top 4 Descriptors Ranked by Chi-square (χ^2).

Model	AUC	CA
Naïve Bayes	0.799	**0.811**
Neural Network	0.751	0.757
Tree	0.724	0.730
Random Forest	0.769	0.622

This time, the order of the algorithms based on their CA scores is as follows: NB > NN > DT > RF. The order is similar to the one from the first experiment, without feature selection, except that NB and RF swapped places, and now the NB achieved a higher AC score. Notably, NB demonstrated an improvement of nearly 11%. This can be viewed as an improvement in adequately placing 8 students out of 10 (instead of approximately 3 out of 10). This gain of accurately placing one more student, combined with NB's reputation for simplicity and minimal computational effort (Pajila et al., 2023), makes it a promising, well-suited algorithm for classification tasks, particularly in the context of DevEd.

7 CONCLUSIONS AND FUTURE WORK

This study evaluated the adequacy of machine learning (ML) based systems, particularly ACCUPLACER, in comparison to human rater classifications within a DevEd placement context. A corpus of 100 essays was assessed using ACCUPLACER'S 6 linguistic descriptors, with two—*Mechanical Conventions* and *Organization and Structure*—linguistically enhanced with quantifiable criteria, to improve inter-rater reliability. Human raters also classified the essays into two DevEd levels (Levels 1 and 2), and these classifications were compared to those assigned by ACCUPLACER.

Significant differences between human and automated classifications were noted, with ACCUPLACER presenting a 22% misplacement rate, compared to the human ratings. This result merits careful consideration. As with any human intelligence task (HIT), discrepancies in both the application of the 6 descriptors and the level assignments were carefully analyzed and resolved, leading to the production of a gold standard for classification—an important contribution that provides a curated dataset for future machine-learning modeling. This gold standard was subsequently used to mimic ACCUPLACER's task using ML algorithms, in which Random Forest achieved a classification accuracy (CA) of 0.757, comparable to a CA baseline of 0.727 obtained with a broader set of automatically extracted linguistic features and used in a prior study (Da Corte and Baptista, 2024c).

Despite ACCUPLACER'S constructs being abstract and challenging for human raters, the results demonstrate that high accuracy can still be achieved when these constructs are enhanced with human input. This was evident in the second experiment with the Naïve Bayes algorithm, which showed a performance improvement of nearly 11% (CA = 0.811). This translates to correctly placing approximately 8 out of 10 students in DevEd courses (versus approximately 3 out of 10). The ranking of descriptors using the Chi-square (χ^2) method further emphasized the importance of refining key features like *Development and Support* and *Organization and Structure*, which ranked highest.

Consequently, this study proposes to focus on refining (and only using) four key descriptors (instead of six) —*Development and Support*, *Organization and Structure*, *Mechanical Conventions*, and *Sentence Variety and Style*—by incorporating more precise linguistic features as criteria. The enhancements will include (i) *Development and Support*, incorporating keywords related to argumentation with reasons and examples to better evaluate how effectively a text presents and supports ideas; (ii) *Organization and Structure* more precisely described and measured with features like word omission, pronoun alternation, and enhanced paragraph composition criteria; (iii) *Mechanical Conventions*, which will identify orthographic features such as contractions, word boundary splits, and punctuation misuse; and (iv) *Sentence Variety and Style*, expanded to include grammatical and lexical-semantic features like word repetition, subject-verb agreement, word precision, and multiword expressions (MWE), which have been used to assess proficiency (Arnold et al., 2018).

These refinements will be tested on a larger corpus, with plans to incorporate 600 additional text samples (from the academic year 2023-2024) that will soon be available, including College Level data previously unavailable. At this stage, a two-step classification procedure is envisaged: (i) College/non-College, followed by (ii) DevEd Level-1/Level 2. Additionally, Large Language Models (LLMs), such as Generative Pre-trained Transformer (GPT), will be leveraged to see how feasible it will be to further align textual features with these refined descriptors and generate explanations for why certain essays meet or fail to meet specific criteria. This could potentially help resolve inter-rater conflicts and improve the overall classification process for DevEd placements, offering critical insights into both writing proficiency and curriculum design. Most importantly, these insights could be used to more effectively prepare and support students in their academic programs.

ACKNOWLEDGMENTS

This work was supported by Portuguese national funds through FCT (Reference: UIDB/50021/2020, DOI: 10.54499/UIDB/50021/2020) and by the European Commission (Project: iRead4Skills, Grant number: 1010094837, Topic: HORIZON-CL2-2022-TRANSFORMATIONS-01-07, DOI: 10.3030/101094837).

We also extend our profound gratitude to the dedicated annotators who participated in this task and the IT team whose expertise made the systematic analysis of the linguistic features presented in this paper possible. Their meticulous work and innovative approach have been instrumental in advancing our research.

REFERENCES

Arnold, T., Ballier, N., Gaillat, T., and Lissón, P. (2018). Predicting CEFRL levels in learner English on the basis of metrics and full texts. *arXiv preprint arXiv:1806.11099*.

Barnett, E. A., Kopko, E., Cullinan, D., and Belfield, C. (2020). Who should take college-level courses? Impact findings from an evaluation of a multiple measures assessment strategy. *Center for the Analysis of Postsecondary Readiness*.

Bickerstaff, S., Beal, K., Raufman, J., Lewy, E. B., and Slaughter, A. (2022). Five principles for reforming developmental education: A review of the evidence. *Center for the Analysis of Postsecondary Readiness*, page 1.

Chen, X. and Meurers, D. (2016). CTAP: A Web-Based Tool Supporting Automatic Complexity Analysis. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 113–119, Osaka, Japan. The COLING 2016 Organizing Committee.

Cohen, J. (1988). *Statistical power analysis*. Hillsdale, NJ: Erlbaum.

Da Corte, M. and Baptista, J. (2022). A phraseology approach in developmental education placement. In *Proceedings of Computational and Corpus-based Phraseology, EUROPHRAS 2022, Malaga, Spain*, pages 79–86.

Da Corte, M. and Baptista, J. (2024a). Charting the linguistic landscape of developing writers: An annotation scheme for enhancing native language proficiency. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3046–3056, Torino, Italia. ELRA and ICCL.

Da Corte, M. and Baptista, J. (2024b). Enhancing writing proficiency classification in developmental education: The quest for accuracy. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6134–6143, Torino, Italia. ELRA and ICCL.

Da Corte, M. and Baptista, J. (2024c). Leveraging NLP and machine learning for English (l1) writing assessment in developmental education. In *Proceedings of the 16th International Conference on Computer Supported Education (CSEDU 2024), 2-4 May, 2024, Angers, France*, volume 2, pages 128–140.

Demšar, J., Curk, T., Erjavec, A., Črt Gorup, Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Štajdohar, M., Umek, L., Žagar, L., Žbontar, J., Žitnik, M., and Zupan, B. (2013). Orange: Data Mining Toolbox in Python. *Journal of Machine Learning Research*, 14:2349–2353.

dos Santos, S. and Junior, G. (2024). Opportunities and challenges of AI to support student assessment in computing education: A systematic literature review. *CSEDU (2)*, pages 15–26.

Dowell, N. and Kovanovic, V. (2022). Modeling educational discourse with natural language processing. *Education*, 64:82.

Duch, D., May, M., and George, S. (2024). Empowering students: A reflective learning analytics approach to enhance academic performance. In *16th International Conference on Computer Supported Education (CSEDU 2024)*, pages 385–396. SCITEPRESS-Science and Technology Publications.

Edgecombe, N. and Weiss, M. (2024). Promoting equity in developmental education reform: A conversation with Nikki Edgecombe and Michael Weiss. *Center for the Analysis of Postsecondary Readiness*, page 1.

Fleiss, J. L. and Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

Freelon, D. (2013). Recal OIR: ordinal, interval, and ratio intercoder reliability as a web service. *International Journal of Internet Science*, 8(1):10–16.

Gallardo, K. (2021). *The Importance of Assessment Literacy: Formative and Summative Assessment Instruments and Techniques*, pages 3–25. Springer Singapore, Singapore.

Ganga, E. and Mazzariello, A. (2019). Modernzing college course placement by using multiple measures. *Education Commission of the States*, pages 1–9.

Giordano, J. B., Hassel, H., Heinert, J., and Phillips, C. (2024). *Reaching All Writers: A Pedagogical Guide for Evolving College Writing Classrooms*, chapter Chapter 2, pages 24–62. University Press of Colorado.

Götz, S. and Granger, S. (2024). Learner corpus research for pedagogical purposes: An overview and some research perspectives. *International Journal of Learner Corpus Research*, 10(1):1–38.

Hirokawa, S. (2018). Key attribute for predicting student academic performance. In *Proceedings of the 10th International Conference on Education Technology and Computers*, pages 308–313.

Huang, Z. (2023). An intelligent scoring system for English writing based on artificial intelligence and machine learning. *International Journal of System Assurance Engineering and Management*, pages 1–8.

Hughes, S. and Li, R. (2019). Affordances and limitations of the ACCUPLACER automated writing placement tool. *Assessing Writing*, 41:72–75.

Kopko, E., Brathwaite, J., and Raufman, J. (2022). The next phase of placement reform: Moving toward equity-centered practice. research brief. *Center for the Analysis of Postsecondary Readiness*.

Lattek, S. M., Rieckhoff, L., Völker, G., Langesee, L.-M., and Clauss, A. (2024). Systematization of competence assessment in higher education: Methods and instruments. In *CSEDU (2)*, pages 317–326.

Link, S. and Koltovskaia, S. (2023). *Automated Scoring of Writing*, pages 333–345. Springer International Publishing, Cham.

Matthews, N. and Folivi, F. (2023). Omit needless words: Sentence length perception. *PloS one*, 18(2):e0282146.

Mukaka, M. M. (2012). A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal*, 24(3):69–71.

Pajila, P. B., Sheena, B. G., Gayathri, A., Aswini, J., Nalini, M., et al. (2023). A comprehensive survey on Naive Bayes algorithm: Advantages, limitations and applications. In *2023 4th International Conference on Smart Electronics and Communication (ICOSEC)*, pages 1228–1234. IEEE.

Perelman, L. (2020). The BABEL generator and e-Rater: 21st Century writing constructs and automated essay scoring (AES). *Journal of Writing Assessment*, 13(1).

Sghir, N., Adadi, A., and Lahmer, M. (2023). Recent advances in predictive learning analytics: A decade systematic review (2012–2022). *Education and information technologies*, 28(7):8299–8333.

The College Board (2022). ACCUPLACER Program Manual. (online).

APPENDIX

Gold standard for the 6 linguistic descriptors and skill levels with a corpus of 100 DevEd text samples.
Data presentation:
ID = document ID; **MC** = Mechanical Conventions; **SVS** = Sentence Variety & Style; **DS** = Development & Support; **OS** = Organization & Structure; **PF** = Purpose & Focus; **CT** = Critical Thinking; and **DevEd** = Development Education Level ("1" or "2").
Likert scale: **0** = deficient; **1** = below average; **2** = above average; and **3** = outstanding.

1,0,1,1,1,1,1,1
3,0,1,1,1,1,1,1
4,1,1,0,0,1,1,1
5,0,1,2,1,2,2,1
9,1,1,1,1,2,1,1
12,1,1,1,2,1,1,1
13,2,1,1,1,2,2,1
21,1,2,1,1,2,1,1
23A,0,0,1,0,1,0,1
23B,0,1,2,1,2,1,1
24A,2,2,2,2,2,1,2
24B,1,1,1,0,1,1,1
25,0,1,1,1,1,0,1
26,0,1,1,1,1,1,1
27,1,1,0,0,0,0,1
29A,0,1,1,0,1,0,1
29B,0,1,1,1,1,1,2
38,1,2,2,1,2,2,1
40A,0,1,2,1,2,2,1
40B,0,0,1,0,0,1,1
40C,1,1,1,1,2,2,1
45,2,2,0,1,1,1,1
48,0,1,1,0,1,1,1
49A,1,1,2,2,2,2,2
49B,1,1,1,0,1,1,1
51,1,2,2,1,2,1,1
55,1,1,1,1,1,1,1
56A,1,1,0,0,0,0,1
56B,1,1,1,1,1,1,1
59,2,1,1,1,1,2,1
60,1,1,1,0,1,1,1
61A,1,1,2,2,1,2,2
61B,0,1,1,1,1,1,2
69A,0,1,2,1,1,1,1
69B,1,1,1,1,2,1,1
70,1,0,0,1,1,1,1
73A,1,1,1,1,1,1,2
73B,2,1,1,1,2,1,2
73C,1,1,1,1,1,1,1
74,0,1,1,2,1,1,1
76,0,0,1,1,1,1,1
77,0,1,2,2,2,2,2
78,0,0,1,0,0,1,1
80A,1,2,2,2,2,1,2
80B,1,0,0,1,1,1,1
81,0,1,1,0,1,1,1
82,2,1,1,1,1,1,2
83,1,1,1,0,1,1,1
85,1,1,2,2,2,2,2
90A,1,2,2,2,2,2,2
90B,0,1,1,1,1,1,1
92A,0,1,0,1,0,0,1
92B,1,1,1,0,0,0,1
93A,0,1,2,2,1,2,2
93B,1,1,1,1,1,1,1
94,2,1,1,1,1,1,1
95A,1,1,1,2,2,1,1
95B,0,0,0,0,0,1,1
96,1,1,1,1,1,1,1
99,1,1,1,1,1,1,1
100,2,2,2,2,3,3,2
103,0,1,2,1,1,1,1
104A,1,1,1,1,1,2,1
104B,2,2,1,1,2,2,2
104C,1,2,1,1,1,1,2
106,1,1,2,2,2,2,1
110,1,2,1,1,1,1,1
113,0,2,1,0,1,1,1
115,0,1,1,0,1,1,1
116,1,1,1,1,1,1,1
118,1,1,1,1,1,1,1
119,2,3,2,2,3,2,2
120,2,1,0,1,2,2,1
123,3,2,1,1,1,1,1
124,1,2,1,1,2,1,1
125A,2,2,3,2,3,2,2
125B,1,1,1,1,1,2,1
125C,1,1,1,1,1,0,1
126,1,1,1,1,2,1,1
132,1,1,2,2,1,2,1
134,2,2,3,3,2,3,2
135,1,1,0,1,1,0,1
138,1,1,1,1,1,1,1
140,3,2,1,3,2,1,2
143A,2,2,3,2,2,2,2
143B,1,2,2,2,2,1,1
147,2,2,1,1,1,2,2
150,0,2,1,1,1,2,1
151A,1,2,1,1,2,1,1
151B,1,1,2,1,2,2,1
152,0,1,1,0,1,1,1
163,2,1,1,1,2,2,1
174,2,2,2,2,2,3,2
178,1,2,2,2,2,2,2
180,2,2,2,2,2,2,2
184,2,2,3,2,2,2,2
187A,1,2,2,2,2,2,1
193,2,3,2,2,3,2,2
198,2,1,1,2,2,1,2
199,3,2,2,1,3,1,1