

# Measuring Writing Proficiency Through Keystroke Process Data: A Cross-School Generalizability Study

Damilola Babalola<sup>1</sup> Naga Buddarapu<sup>1</sup> Caleb Scott<sup>1</sup> Piotr Mitros<sup>2</sup> Paul Deane<sup>2</sup> Collin Lynch<sup>1</sup>

<sup>1</sup> North Carolina State University

<sup>2</sup> Educational Testing Service

{djbabalo,vbuddar,cwscott3,cflynch}@ncsu.edu

{pmitros,pdeane}@ets.org

## Abstract

Keystroke logging offers a scalable way to observe writing as it unfolds, but its value for educational measurement depends on whether process features are valid and stable across schools, grades, and writing contexts. We analyze 8,576 authentic classroom documents from 423 seventh- and eighth-grade students in two U.S. middle schools. Twenty-four theory-motivated keystroke measures were organized into fluency, revision, complexity, and engagement families and expanded through distributional summaries of timing measures. We examine associations between process features and incoming state English Language Arts (ELA) proficiency, the interpretability and stability of latent factor structures, and the consistency of gender-linked differences across schools and writing contexts. Results show that several process features, particularly those related to engagement and timing, were associated with ELA proficiency, with the strongest and most consistent patterns emerging in Grade 8. These associations showed moderate replication across schools and writing contexts. Factor analyses identified recurring dimensions related to engagement, revision, timing variability, and planning, although their structure was not fully stable across cohorts. Gender-linked differences were also inconsistent across grades and schools. Overall, the findings support keystroke measures as promising process signals for AI-enabled formative assessment while highlighting the need for local validation, adequate sampling, and fairness checks before deployment.

## 1 Introduction

Writing proficiency is expressed not only in the quality of a finished text but also in the processes through which that text is planned, produced, reviewed, and revised. Cognitive models describe writing as a recursive, self-regulated activity in which writers allocate limited attention and

working-memory resources across planning, translating, and reviewing (Flower and Hayes, 1981; Graham et al., 2017; Zimmerman and Risemberg, 1997). Traditional product-focused assessments observe the outcome of these processes, emphasizing text quality, coherence, and surface-level correctness, but provide limited evidence about how students reached it. With advances in educational data mining, keystroke logging now enables the fine-grained observation of writing in real-time, capturing insertions, deletions, pauses, and navigational behaviors (Leijten and Van Waes, 2013; Baaijen et al., 2012).

This process evidence is increasingly relevant to artificial intelligence (AI) in educational measurement. Automated writing evaluation has historically relied primarily on features extracted from the final text, but more recent studies have used keystroke traces in machine-learning models to predict writing quality, identify students who may need support, distinguish human from AI-generated text, and differentiate planning strategies across student clusters (Conijn et al., 2022; Litman et al., 2022; Jiang et al., 2024). More recent measurement research has also modeled raw keystroke sequences with large language model architectures to infer stable characteristics of writers (Zhang et al., 2025). These applications make keystroke data attractive as an additional behavioral channel for AI-enabled assessment, but they also raise a basic measurement question: before process measures are used for prediction, feedback, or automated decision-making, are the signals sufficiently stable and generalizable across educational contexts?

Existing keystroke studies provide evidence that timing, fluency, revision, and time-on-task measures can relate to writing quality, but much of that evidence comes from controlled or time-limited tasks, adult or postsecondary writers, or a single assessment context (Leijten and Van Waes, 2013;

Conijn et al., 2022, 2019). Authentic middle-school writing is less standardized: students write across subjects, assignments, teachers, and time spans, and the meaning of a pause or revision may vary with task demands (Conijn et al., 2019). A signal that is predictive in one cohort may therefore fail to transfer to another. This is especially important for classroom-embedded AI systems, where instability can be mistaken for student differences and where demographic patterns can become embedded in automated interpretations.

We address this issue using a full academic year of naturally occurring writing from two U.S. middle schools. The study treats cross-context replication as a measurement requirement rather than as an afterthought. RQ1, our primary focus, asks which keystroke-derived behaviors are associated with students' incoming ELA proficiency and whether those associations replicate across schools, grades, and writing contexts. RQ2 asks whether the features form interpretable latent dimensions and whether those structures generalize. RQ3 provides an exploratory fairness-oriented check of whether gender-linked feature differences are stable or context-bound. Together, the analyses evaluate whether process features are plausible candidates for classroom-embedded, AI-supported formative measurement, where process evidence may complement established approaches to formative writing assessment (Graham et al., 2015).

## 2 Background

### 2.1 From Keystrokes to Writing Constructs

Keystroke logging has become instrumental in identifying writing strategies and understanding cognitive processes by making the process underlying the written product observable (Deane et al., 2026). To interpret keystroke data meaningfully, low-level behavioral metrics must be linked to higher-order cognitive constructs (Deane et al., 2008; Odendahl and Deane, 2018). The present study organizes its features into four theoretical families grounded in established models of writing cognition:

- **Fluency** reflects the efficiency and rhythm with which ideas are translated into text, including insertion activity and inter-keystroke intervals within and between words (Van Waes and Leijten, 2015).
- **Revision** captures self-monitoring and correc-

tion through edits, backspaces, and deletion sequences, consistent with models of writing as a self-regulated process (Baaijen et al., 2012; Graham et al., 2017; Zimmerman and Risemberg, 1997).

- **Complexity** captures timing and behavior at clause and punctuation boundaries, where sentence-level and rhetorical planning may become visible (Flower and Hayes, 1981; Deane et al., 2008).
- **Engagement** captures sustained allocation of attention through active writing, reading, navigation, and menu time (Conijn et al., 2019).

The feature families are theoretical reference frames rather than claims that a single keystroke event uniquely identifies a cognitive state. Pauses at different textual locations can reflect different underlying processes, including lexical processing and higher-level planning, making one-to-one mappings between pause behavior and cognition difficult (Conijn et al., 2019; Révész et al., 2026). This approach also motivates the latent-structure analysis in RQ2: if multiple observed features repeatedly cluster in interpretable ways, that convergence provides additional evidence about construct representation, while instability cautions against universal feature-to-construct mappings.

### 2.2 Keystroke Data in AI-Enabled Assessment

Process data are already being incorporated into AI and machine-learning approaches to writing assessment. Conijn et al. (2022) trained regression and classification models on keystroke features to test early prediction of writing quality and identification of writers who might need support. Jiang et al. (2024) used keystroke behavior patterns in machine-learning detectors of nonauthentic writing in a large-scale assessment and explicitly evaluated fairness across demographic groups. Zhang et al. (2025) described both theory-driven feature modeling and an approach in which raw keystroke logs were input to large language models to infer stable writer characteristics. These studies show that keystroke traces are not merely descriptive research data; they are becoming inputs to AI-supported measurement systems.

More broadly, unobtrusively collected behavioral traces have also been used for “stealth” assessment, in which learner performance is in-

ferred from activity generated during authentic tasks rather than from a separate test (Shute et al., 2016). That trajectory strengthens the need for validation studies that test portability before deployment. Keystroke features can be sensitive to task demands (Conijn et al., 2019), meaning that a signal that is statistically useful in one dataset may not have the same distribution or interpretation elsewhere. Our emphasis on cross-school, cross-grade, and cross-context stability therefore complements predictive-model research: it asks whether the behavioral signals themselves behave consistently enough to support downstream AI use.

### 3 Data and Methods

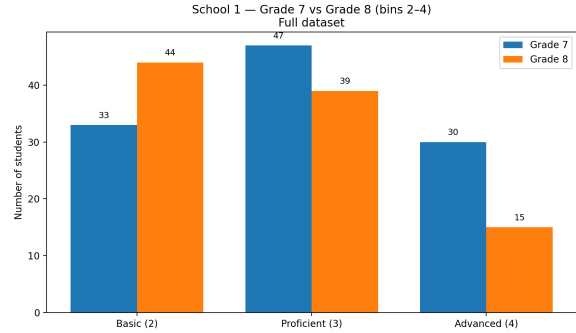
#### 3.1 Participants, Writing Data, and Proficiency Measure

Data were collected during the 2023–2024 academic year from students in Grades 7 and 8 at two U.S. middle schools participating in a research project. School 1 is a large urban school in Utah (857 students; 42% White, 42% Hispanic, 8% multiracial, 3.5% African American, 3% Asian), and School 2 is a suburban school in Pennsylvania (713 students; 75% White, 7% Hispanic, 7% Asian American, 4.2% African American). A classroom-embedded Chrome extension captured Google Docs writing activity, including insertions, deletions, pauses, navigation, and reading-related events. The logs span ELA assignments (e.g., essays, narratives, text-dependent analyses) and non-ELA writing (e.g., lab reports, notes). The non-ELA category is important because students engage in substantial writing activities across subjects, while explicit writing instruction typically occurs in ELA. The study was conducted under Institutional Review Board approval with pseudonymized data handling.

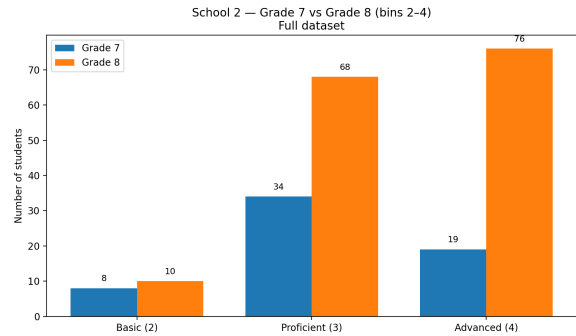
The initial collection contained 9,831 documents from 544 students. Documents shorter than 50 words were excluded. Students’ prior-year (2022–2023) state ELA proficiency scores served as the reference performance measure. Although assessments differ by state, scores were harmonized into four bins (1 = Below Basic, 2 = Basic, 3 = Proficient, 4 = Advanced). Because School 2 contained no Below-Basic cases and School 1 contained few, bin 1 was excluded, and the remaining three levels were used to ensure comparability across schools. The final corpus consisted of 8,576 documents (School 1: 3,641; School 2:

| School   | Context | Students |     | Documents |       |
|----------|---------|----------|-----|-----------|-------|
|          |         | G7       | G8  | G7        | G8    |
| School 1 | Full    | 110      | 98  | 1,840     | 1,801 |
| School 1 | ELA     | 102      | 93  | 510       | 417   |
| School 2 | Full    | 61       | 154 | 989       | 3,946 |
| School 2 | ELA     | 55       | 153 | 170       | 806   |

Table 1: Analytic sample by school, context, grade, and document count.



(a) School 1



(b) School 2

Figure 1: Distribution of state ELA proficiency levels by grade and school.

4,935) from 423 students. We analyze two writing contexts. The Full dataset contains all eligible classroom writing while the ELA subset focuses on extended ELA writing identified through assignment-title keywords (e.g., Essay, Narrative, Argument, Analysis, TDA) and minimum thresholds of 200 words and 50 insertion events. Comparing the two contexts tests whether associations observed in ELA writing also hold across the broader scope of classroom writing. Student demographics (gender, grade, class period, English Learner status) were obtained from school rosters. Table 1 summarizes sample sizes by school, grade, and dataset. Figures 1a and 1b show the distribution of both schools for the full dataset.

| Dimension  | Keystroke measures                                                                                                                                                                                                                                                                                                                    |
|------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fluency    | Number of inserts; in-word inserts; between-word inserts; word-start inserts; IKI between inserts (interval between successive insertions); IKI within-word (interval between letters); IKI between words (interval between words); IKI before-word (interval preceding word onset); IKI at word-start (interval at word initiation). |
| Revision   | Before-word inserts; total edits; starting backspaces (at word start); continuing backspaces (sequential deletions across words); IKI starting backspace (interval before deletion onset); IKI continuing backspace (interval during deletion sequences); edit duration.                                                              |
| Complexity | Clause-break inserts (marking syntactic boundaries); minor punctuation inserts (within-sentence punctuation); IKI at clause-break (interval at clause boundaries); IKI at minor punctuation (interval surrounding within-sentence punctuation).                                                                                       |
| Engagement | Overall menu time; overall reading time (inactive but focused); overall navigation time (cursor, scroll, and arrow-key movement); overall active writing time.                                                                                                                                                                        |

Table 2: Theory-motivated keystroke feature families used in the analysis.

### 3.2 Feature Extraction

The keystroke logs were transformed into quantitative indicators of writing behavior from time-stamped keypress, deletion, pause, and navigation events. Following the framework in Section 2.1, the 24 extracted measures were grouped into fluency, revision, complexity, and engagement dimensions (Table 2). Ten measures were frequency based and fourteen were time based. Inter-keystroke interval (IKI) denotes the elapsed time between successive keystroke or editing events and captures temporal aspects of writing behavior.

To reduce the influence of heavy-tailed timing distributions, time-based measures were transformed using  $\log(1 + x)$ , and pauses exceeding 120 seconds were truncated as likely off-task intervals. For timing measures other than overall writing, reading, navigation, and menu time, we calculated the mean, median, standard deviation, and variance, yielding 54 analytic variables (Leijten and Van Waes, 2013; Baaijen et al., 2012; Conijn et al., 2022, 2019; Van Waes and Leijten, 2015). In the variable names for the analyses, *mean*, *median*, *std*, and *var* denote the distributional summaries, while *n* denotes a count.

### 3.3 Analyses

- **RQ1:** Within each School  $\times$  Grade  $\times$  Context cohort, we computed Spearman rank correlations between each feature and the ordered incoming ELA proficiency level. Benjamini-Hochberg false discovery rate (FDR) correction (Benjamini and Hochberg, 1995) was applied at  $\alpha = .10$ . To test generalizability, sets of significant features were compared using Jaccard similarity (Jaccard, 1901) across (a) ELA versus Full contexts within school and grade, (b) Grade 7 versus Grade 8 within school and context, and (c) School 1 vs School 2 within grade and context. Directional agreement was calculated for shared features, and correlation analyses used students with valid proficiency scores and available feature data; consequently, analytic *N* may be slightly smaller than the cohort counts reported in Table 1.
- **RQ2:** Exploratory factor analysis (EFA) was conducted separately by cohort using standardized features, principal-axis factoring, and Promax rotation. Extreme values were winsorized at the 1% tails. Cohorts required  $N \geq 40$ , and sampling adequacy was assessed using the Kaiser-Meyer-Olkin (KMO) statistic (Kaiser, 1960), with values below .60 treated as marginal. Factorability was also assessed using Bartlett’s test of sphericity (Bartlett, 1950), and loadings with  $|\lambda| \geq .30$  were treated as salient. Factor retention used parallel analysis (Horn, 1965) together with the eigenvalue-greater-than-one criterion, conservatively retaining the smaller estimate. Tucker’s congruence coefficient (Lorenzo-Seva and ten Berge, 2006) quantified similarity between factor-loading structures across contexts, grades, and schools. Because several cohorts are modest in size relative to the number of variables, low congruence is interpreted as potentially reflecting both genuine context sensitivity and estimation instability (MacCallum et al., 1999).
- **RQ3:** Gender comparisons were exploratory and limited to binary roster labels (Female/Male), the only consistently available demographic field across sites. Two-sided Mann-Whitney U tests were applied within cohorts with FDR correction at  $\alpha = .10$ ,

and Cliff’s delta ( $\delta$ ) summarized effect magnitude and direction (Cliff, 1993). Analyses required at least 10 students per gender group and 30 students per cohort. Jaccard overlap and directional consistency assessed replication across contexts, grades, and schools.

## 4 Results

### 4.1 RQ1: Proficiency Associations and Generalizability

The clearest proficiency associations emerged in Grade 8 (detailed feature-level results are reported in Appendix A.1). In School 1 ELA writing, overall reading time ( $\rho = .35$ ), between-word timing variability ( $\rho = .34$ ), and clause-break timing variability ( $\rho = .30$ ) were positively associated with incoming ELA proficiency, while longer mean clause-boundary pauses were negatively associated ( $\rho \approx -.24$ ). The Full dataset showed a similar pattern, with reading, active-writing, and between-word activity among the strongest signals.

School 2 showed an even clearer Grade 8 profile: navigation, reading, and active-writing time correlated positively with proficiency ( $\rho \approx .35-.41$ ), whereas mean and median clause-boundary timing correlated negatively ( $\rho \approx -.40$ ). Together, these results suggest that students with higher incoming ELA proficiency were characterized not simply by greater text production, but by sustained engagement and distinct temporal patterns, including shorter pauses at linguistic boundaries.

Neither school’s Grade 7 ELA subset produced significant feature–proficiency associations after FDR correction. In the Full datasets, School 1 Grade 7 showed three associations involving active writing, menu time, and between-word insertions, whereas School 2 Grade 7 showed only a navigation-time association. Differences in cohort size, score distributions, and range restriction prevent interpreting this pattern as developmental change.

Replication analyses nevertheless reinforced the Grade 8 result (Table 3). ELA–Full overlap was moderate in both schools ( $J = .53$  and  $.58$ ) with perfect directional agreement, while cross-school Grade 8 overlap reached  $J = .36$  in ELA and  $J = .58$  in the Full dataset. Thus, the strongest Grade 8 fluency and engagement signals generalized beyond a single school or writing context.

| Comparison   | Cohort   | Jacc. | Overlap | Sign |
|--------------|----------|-------|---------|------|
| ELA vs. Full | S1, G8   | 0.53  | 8       | 1.00 |
| ELA vs. Full | S2, G8   | 0.58  | 15      | 1.00 |
| S1 vs. S2    | ELA, G8  | 0.36  | 9       | 0.89 |
| S1 vs. S2    | Full, G8 | 0.58  | 5       | 1.00 |
| G7 vs. G8    | S1, Full | 0.27  | 3       | 1.00 |
| G7 vs. G8    | S2, Full | 0.05  | 1       | 1.00 |

Table 3: Stability of significant RQ1 feature–proficiency associations. Overlap is the number of shared significant features; Sign is the proportion with matching direction.

### 4.2 RQ2: Latent Behavioral Structure

Exploratory factor analysis yielded five to eight factors across cohorts. Despite variation in exact composition, several recurring behavioral dimensions emerged. A global-engagement factor commonly combined active-writing, reading, and navigation time. Revision-related factors separated rapid local revision from more sustained editing, with insertion and backspace intervals reflecting revision fluency and longer edit durations capturing deliberate multi-step refinement. Timing variability formed another recurring dimension, while pre-word pauses and clause- or punctuation-level timing reflected planning and linguistic rhythm. These planning and clause-level patterns were more prominent in Grade 8.

The Grade 8 ELA solutions illustrate both the recurring behavioral dimensions and differences in their precise composition across schools. In both schools, active-writing, navigation, reading, and menu time clustered within a global-engagement factor, while continuing-backspace and edit-duration measures formed a sustained-revision factor. The ordering and composition of the remaining factors differed across schools (Appendix A.2).

Bartlett’s tests of sphericity were significant across all analyzed cohorts, supporting factorability. Sampling adequacy ranged from marginal to acceptable ( $KMO = .52-.78$ ), with the weakest values occurring in School 2 Grade 7. Although these recurring dimensions were interpretable, their exact factor composition was not stable across cohorts. Cross-context congruence was negligible in Grade 7 but higher in Grade 8, particularly for School 1 ( $\phi = .66$ ). Cross-grade and cross-school congruence remained below conventional similarity thresholds.

Thus, the data support recurring families of en-

| School   | Grade | ELA KMO  | Full KMO | Congr. |
|----------|-------|----------|----------|--------|
| School 1 | G7    | .782 (6) | .757 (8) | .01    |
|          | G8    | .767 (6) | .764 (6) | .66    |
| School 2 | G7    | .520 (5) | .609 (5) | .07    |
|          | G8    | .725 (6) | .721 (7) | .37    |

Table 4: RQ2 sampling adequacy and cross-context factor stability. KMO values are followed in parentheses by the number of retained factors. Congr. denotes ELA–Full Tucker congruence.

gagement, revision, timing variability, and planning behavior, but not a single invariant latent structure. Given the modest size of several cohorts, this instability may reflect both genuine context sensitivity and factor-estimation uncertainty (MacCallum et al., 1999).

### 4.3 RQ3: Exploratory Gender Differences

Gender-linked differences were less stable than proficiency associations. In School 1 Grade 7, female students showed greater between-word insertion activity ( $\delta = -.57$ ), navigation ( $\delta = -.52$ ), active-writing activity ( $\delta = -.51$ ), and reading activity ( $\delta = -.48$ ), whereas male students showed greater clause-boundary timing variability ( $\delta = .52$ ). These effects replicated across ELA and Full writing contexts. In contrast, no significant gender differences were observed for School 1 Grade 8.

School 2 showed a different pattern. Grade 7 ELA produced no significant differences, while the Full dataset showed only a small number of male-dominant word-start timing effects. In Grade 8, female students showed somewhat greater engagement, whereas male students showed longer or more variable clause- and word-start pauses. Detailed feature-level gender comparisons are reported in Appendix A.3.

Within-cohort replication could therefore be moderate to strong, but cross-grade and cross-school persistence was minimal or absent (Table 5). The results do not support gender as a stable explanation of keystroke behavior; instead, observed differences appear contingent on cohort and writing context.

## 5 Discussion

This study examined whether keystroke-derived writing-process measures remain meaningful across schools, grades, and writing contexts. The strongest evidence emerged in Grade 8, where higher incoming ELA proficiency was consistently

| Comparison   | Cohort    | Jacc. | Overlap | Interpretation                     |
|--------------|-----------|-------|---------|------------------------------------|
| ELA vs. Full | S1, G7    | .57   | 12      | Strong within-cohort replication   |
| ELA vs. Full | S2, G8    | .50   | 6       | Moderate within-cohort replication |
| G7 vs. G8    | All pairs | .00   | 0       | No cross-grade persistence         |
| S1 vs. S2    | Full, G7  | .11   | 2       | Minimal cross-school replication   |

Table 5: RQ3 stability of significant gender-linked feature differences. S1 and S2 denote Schools 1 and 2; G7 and G8 denote Grades 7 and 8.

associated with greater reading, navigation, and active-writing activity and with shorter pauses at clause boundaries. Replication across schools and across ELA and broader classroom writing suggests that these relationships are not confined to a single institutional or task context. These findings are broadly consistent with prior work showing that keystroke-derived timing, fluency, revision, and engagement measures contain information related to writing performance (Conijn et al., 2022; Van Waes and Leijten, 2015). At the same time, prior evidence that keystroke features are sensitive to writing-task demands (Conijn et al., 2019) makes the cross-context replication observed here particularly important.

At the same time, the findings caution against treating all process measures as portable indicators. Grade 7 associations were sparse, factor structures were not invariant, and gender-linked differences did not replicate consistently. These results matter for AI-enabled educational measurement because predictive usefulness within one cohort does not by itself establish the validity of an interpretation or its transferability to other populations and contexts (American Educational Research Association et al., 2014). A model may exploit local timing, task, or demographic patterns while performing poorly or differently elsewhere.

The results therefore support a staged approach to using keystroke data in AI systems. Candidate features should first be theoretically motivated and locally validated, then tested for replication across relevant contexts and subgroups. Predictive models should be evaluated not only for accuracy but also for stability and subgroup performance, particularly because automated scoring models can reproduce measurement and algorithmic biases through their features and modeling decisions (Johnson et al., 2022; Jiang et al., 2024).

Process-based feedback should likewise remain tied to observable behavior rather than treating isolated keystroke events as direct evidence of internal cognitive states.

The recurring fluency, revision, engagement, and planning-related patterns provide some support for the interpretability of the feature design, but the unstable factor structures suggest that these categories are better treated as organizing frameworks than fixed latent scales. This interpretation is consistent with prior cautions that observable keystroke behaviors do not map uniquely onto higher-order writing processes (Deane et al., 2008; Conijn et al., 2019). For downstream AI applications, keystroke features may therefore be most useful as context-sensitive evidence sources rather than universal proxies for proficiency.

Among the features examined, reading, navigation, active-writing time, and clause-boundary timing appear most promising because their relationships with proficiency replicated across Grade 8 schools and writing contexts. By contrast, features or factor structures observed in only one cohort should be treated as locally informative until they replicate elsewhere. This distinction is important as keystroke data are increasingly used for writing-quality prediction, detection of nonauthentic text, and other AI-supported assessment tasks (Conijn et al., 2022; Jiang et al., 2024; Zhang et al., 2025; Deane et al., 2026).

## 6 Limitations and Future Work

First, excluding Below-Basic students was necessary to harmonize proficiency categories across states but restricted the criterion range, which can reduce observed correlations (Mendoza and Mumford, 1987). The findings therefore apply only to students represented in the Basic, Proficient, and Advanced categories.

Second, the Grade 7/Grade 8 comparison is cross-sectional. Differences in sample size, score distributions, assignment mix, or cohort composition may explain the stronger Grade 8 associations; longitudinal data are needed before interpreting this pattern developmentally.

Third, some EFA cohorts were small relative to the 54 analytic variables, and one cohort had  $KMO=.52$ . Low Tucker congruence may therefore reflect estimation instability as well as genuine contextual variation (MacCallum et al., 1999). Larger samples and confirmatory or resampling-

based analyses are needed before defining reusable latent scales.

Finally, gender analyses were limited to binary roster labels and do not constitute comprehensive fairness evidence; fairness in educational measurement concerns the validity and comparability of score interpretations and uses across relevant populations, rather than any single demographic comparison (American Educational Research Association et al., 2014; Johnson et al., 2022). Future work should examine other relevant student characteristics where ethically and legally appropriate.

## 7 Conclusion

Keystroke process data provide information about writing behavior that is not visible in final text alone, but their value depends on validation across contexts. Across two middle schools and thousands of classroom documents, Grade 8 proficiency associations involving engagement and temporal fluency replicated across schools and writing contexts, whereas Grade 7 associations were sparse, latent structures were unstable, and gender-linked patterns were context dependent. These findings support cautious use of keystroke measures in AI-enabled formative assessment and reinforce a central requirement: behavioral process signals should demonstrate validity, generalizability, and fairness before being used for automated inference or feedback.

## References

- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 2014. *Standards for Educational and Psychological Testing*. American Educational Research Association, Washington, DC.
- Veerle M. Baaijen, David Galbraith, and Kees de Glopper. 2012. Keystroke analysis: Reflections on procedures and measures. *Written Communication*, 29(3):246–277.
- Maurice S Bartlett. 1950. Tests of significance in factor analysis. *British Journal of Statistical Psychology*, 3(2):77–85.
- Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300.
- Norman Cliff. 1993. Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, 114(3):494–509.

- Rianne Conijn, Christine Cook, Menno van Zaanen, and Luuk Van Waes. 2022. [Early prediction of writing quality using keystroke logging](#). *International Journal of Artificial Intelligence in Education*, 32:835–866.
- Rianne Conijn, Jens Roeser, and Menno van Zaanen. 2019. Understanding the keystroke log: The effect of writing task on keystroke features. *Reading and Writing*, 32(9):2353–2374.
- Paul Deane, Neil Odendahl, Thomas Quinlan, Mary Fowles, Cindy Welsh, and Jennifer Bivens-Tatum. 2008. Cognitive models of writing: Writing proficiency as a complex integrated skill. *ETS Research Report Series*, 2008(2):i–36.
- Paul Deane, Mo Zhang, Jiangang Hao, and Chen Li. 2026. Using keystroke dynamics to detect nonoriginal text. *Journal of Educational Measurement*, 63(1):e12431.
- Linda Flower and John R. Hayes. 1981. A cognitive process theory of writing. *College Composition and Communication*, 32(4):365–387.
- Steve Graham, Karen R. Harris, Charles MacArthur, and Tanya Santangelo. 2017. Self-regulation and writing. In *Handbook of Self-Regulation of Learning and Performance*, pages 138–152. Routledge.
- Steve Graham, Michael Hebert, and Karen R. Harris. 2015. Formative assessment and writing: A meta-analysis. *The Elementary School Journal*, 115(4):523–547.
- John L. Horn. 1965. A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2):179–185.
- Paul Jaccard. 1901. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:547–579.
- Yang Jiang, Mo Zhang, Jiangang Hao, Paul Deane, and Chen Li. 2024. Using keystroke behavior patterns to detect nonauthentic texts in writing assessments: Evaluating the fairness of predictive models. *Journal of Educational Measurement*, 61(4):571–594.
- Matthew S. Johnson, Xiang Liu, and Daniel F. McCaffrey. 2022. [Psychometric methods to evaluate measurement and algorithmic bias in automated scoring](#). *Journal of Educational Measurement*, 59(3):338–361.
- Henry F. Kaiser. 1960. The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20(1):141–151.
- Mariëlle Leijten and Luuk Van Waes. 2013. Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes. *Written Communication*, 30(3):358–392.
- Diane Litman, Tazin Afrin, Omid Kashefi, Christopher Olshefski, Amanda Godley, and Rebecca Hwa. 2022. An automated writing evaluation system for supporting self-monitored revising. In *International Conference on Artificial Intelligence in Education*, pages 581–587. Springer.
- Urbano Lorenzo-Seva and Jos M. F. ten Berge. 2006. Tucker’s congruence coefficient as a meaningful index of factor similarity. *Methodology*, 2(2):57–64.
- Robert C. MacCallum, Keith F. Widaman, Shaobo Zhang, and Sehee Hong. 1999. Sample size in factor analysis. *Psychological Methods*, 4(1):84–99.
- Jorge L. Mendoza and Michael D. Mumford. 1987. [Corrections for attenuation and range restriction on the predictor](#). *Journal of Educational Statistics*, 12(3):282–293.
- Neil Odendahl and Paul Deane. 2018. Assessing the writing process: A review of current practice. Technical Report ETS-RM-18-10, Educational Testing Service.
- Andrea Révész, Hyeonjeong Jeong, Yuyong Liu, Tsubasa Otsuki, Xin Rong, Ting Zeng, Makoto Takahashi, and Motoaki Sugiura. 2026. [The neural correlates of pausing behaviours in L2 writing](#). *Journal of Second Language Writing*, 72:101303.
- Valerie J. Shute, Lihua Wang, Samuel Greiff, Wei Zhao, and Gregory Moore. 2016. Measuring problem solving skills via stealth assessment in an engaging video game. *Computers in Human Behavior*, 63:106–117.
- Luuk Van Waes and Mariëlle Leijten. 2015. Fluency in writing: A multidimensional perspective on writing fluency applied to L1 and L2. *Computers and Composition*, 38:79–95.
- Mo Zhang, Paul Deane, Andrew Hoang, Hongwen Guo, and Chen Li. 2025. Applications and modeling of keystroke logs in writing assessments. *Educational Measurement: Issues and Practice*, 44(2):5–19.
- Barry J. Zimmerman and Rafael Risemberg. 1997. Becoming a self-regulated writer: A social cognitive perspective. *Contemporary Educational Psychology*, 22(1):73–101.

## A Detailed Results

### A.1 RQ1: Feature–Proficiency Associations

Tables 6–9 report significant feature–proficiency associations after FDR correction. Here,  $G$  denotes grade and  $p_{\text{adj}}$  denotes the FDR-adjusted  $p$ -value.

| G | Feature                        | $\rho$  | $p_{adj}$ |
|---|--------------------------------|---------|-----------|
| 8 | overallReadingTime             | 0.3481  | 0.0242    |
| 8 | stdLogBetweenWordInsertTime    | 0.3366  | 0.0242    |
| 8 | varLogClauseBreakInsertTime    | 0.3024  | 0.0474    |
| 8 | overallMenuTime                | 0.2927  | 0.0474    |
| 8 | nBetweenWordInserts            | 0.2887  | 0.0474    |
| 8 | stdLogMinorPunctInsertTime     | 0.2823  | 0.0480    |
| 8 | stdLogWordStartInsertTime      | 0.2709  | 0.0575    |
| 8 | overallActiveWritingTime       | 0.2504  | 0.0770    |
| 8 | varLogBetweenWordInsertTime    | 0.2497  | 0.0770    |
| 8 | meanLogClauseBreakInsertTime   | -0.2440 | 0.0770    |
| 8 | overallNavigationTime          | 0.2437  | 0.0770    |
| 8 | medianLogClauseBreakInsertTime | -0.2344 | 0.0893    |

Table 6: Significant feature–ELA proficiency correlations for School 1 in the ELA writing context.

| G | Feature                     | $\rho$ | $p_{adj}$ |
|---|-----------------------------|--------|-----------|
| 7 | overallActiveWritingTime    | 0.2931 | 0.0546    |
| 7 | overallMenuTime             | 0.2838 | 0.0546    |
| 7 | nBetweenWordInserts         | 0.2561 | 0.0947    |
| 8 | nBetweenWordInserts         | 0.3851 | 0.0024    |
| 8 | overallReadingTime          | 0.3811 | 0.0024    |
| 8 | overallActiveWritingTime    | 0.3327 | 0.0119    |
| 8 | stdLogBetweenWordInsertTime | 0.2953 | 0.0286    |
| 8 | overallMenuTime             | 0.2938 | 0.0286    |
| 8 | varLogWordStartInsertTime   | 0.2829 | 0.0341    |
| 8 | overallNavigationTime       | 0.2622 | 0.0555    |
| 8 | varLogBetweenWordInsertTime | 0.2438 | 0.0764    |
| 8 | varLogInWordInsertTime      | 0.2423 | 0.0764    |
| 8 | varLogStartingBackspaceTime | 0.2337 | 0.0871    |
| 8 | varLogClauseBreakInsertTime | 0.2283 | 0.0913    |

Table 7: Significant feature–ELA proficiency correlations for School 1 in the Full writing dataset.

| G | Feature                          | $\rho$  | $p_{adj}$ |
|---|----------------------------------|---------|-----------|
| 8 | medianLogClauseBreakInsertTime   | -0.4328 | 0.0000    |
| 8 | overallNavigationTime            | 0.4067  | 0.0000    |
| 8 | meanLogClauseBreakInsertTime     | -0.4007 | 0.0000    |
| 8 | overallReadingTime               | 0.3794  | 0.0000    |
| 8 | overallActiveWritingTime         | 0.3515  | 0.0000    |
| 8 | nBetweenWordInserts              | 0.3334  | 0.0002    |
| 8 | meanLogInWordInsertTime          | -0.2838 | 0.0027    |
| 8 | medianLogWordStartInsertTime     | -0.2709 | 0.0043    |
| 8 | medianLogInWordInsertTime        | -0.2629 | 0.0055    |
| 8 | varLogStartingBackspaceTime      | 0.2518  | 0.0080    |
| 8 | varLogContinuingBackspaceTime    | 0.2417  | 0.0111    |
| 8 | meanLogWordStartInsertTime       | -0.2282 | 0.0171    |
| 8 | overallMenuTime                  | 0.2267  | 0.0171    |
| 8 | medianLogEditTime                | -0.2187 | 0.0216    |
| 8 | varLogEditTime                   | 0.2123  | 0.0255    |
| 8 | meanLogContinuingBackspaceTime   | -0.2096 | 0.0264    |
| 8 | medianLogContinuingBackspaceTime | 0.2004  | 0.0344    |
| 8 | medianLogStartingBackspaceTime   | -0.1983 | 0.0350    |
| 8 | meanLogEditTime                  | -0.1832 | 0.0546    |
| 8 | meanLogStartingBackspaceTime     | -0.1814 | 0.0551    |
| 8 | varLogClauseBreakInsertTime      | 0.1774  | 0.0594    |
| 8 | stdLogMinorPunctInsertTime       | -0.1703 | 0.0705    |

Table 8: Significant feature–ELA proficiency correlations for School 2 in the ELA writing context.

| G | Feature                        | $\rho$  | $p_{adj}$ |
|---|--------------------------------|---------|-----------|
| 7 | overallNavigationTime          | 0.3974  | 0.0622    |
| 8 | overallReadingTime             | 0.4735  | 0.0000    |
| 8 | nBetweenWordInserts            | 0.4093  | 0.0000    |
| 8 | overallNavigationTime          | 0.3961  | 0.0000    |
| 8 | overallActiveWritingTime       | 0.3858  | 0.0000    |
| 8 | medianLogClauseBreakInsertTime | -0.3825 | 0.0000    |
| 8 | medianLogWordStartInsertTime   | -0.3414 | 0.0001    |
| 8 | meanLogClauseBreakInsertTime   | -0.3399 | 0.0001    |
| 8 | meanLogWordStartInsertTime     | -0.3026 | 0.0006    |
| 8 | meanLogInWordInsertTime        | -0.3016 | 0.0007    |
| 8 | medianLogInWordInsertTime      | -0.2975 | 0.0007    |
| 8 | medianLogStartingBackspaceTime | -0.2660 | 0.0032    |
| 8 | stdLogBeforeWordInsertTime     | -0.2629 | 0.0034    |
| 8 | varLogContinuingBackspaceTime  | 0.2608  | 0.0034    |
| 8 | meanLogStartingBackspaceTime   | -0.2507 | 0.0050    |
| 8 | overallMenuTime                | 0.2407  | 0.0072    |
| 8 | stdLogInWordInsertTime         | -0.1920 | 0.0437    |
| 8 | stdLogMinorPunctInsertTime     | -0.1840 | 0.0525    |
| 8 | stdLogEditTime                 | -0.1831 | 0.0525    |
| 8 | medianLogBetweenWordInsertTime | -0.1619 | 0.0969    |

Table 9: Significant feature–ELA proficiency correlations for School 2 in the Full writing dataset.

## A.2 RQ2: Illustrative Factor Solutions

The following tables present the Grade 8 ELA factor solutions for both schools as illustrative examples of the recurring behavioral dimensions and variation in their composition across schools.

| Factor | Dominant features                                                                                                                                                                                                                 | Theme                  |
|--------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------|
| F1     | overallNavigationTime,<br>overallActiveWritingTime,<br>overallReadingTime,<br>overallMenuTime,<br>varLogMinorPunctInsertTime,<br>varLogBetweenWordInsertTime,<br>varLogStartingBackspaceTime,<br>varLogInWordInsertTime           | Global engagement time |
| F2     | medianLogInWordInsertTime,<br>meanLogStartingBackspaceTime,<br>medianLogStartingBackspaceTime,<br>medianLogBetweenWordInsertTime,<br>meanLogBetweenWordInsertTime,<br>meanLogWordStartInsertTime,<br>medianLogWordStartInsertTime | Quick revision         |
| F3     | meanLogContinuingBackspaceTime,<br>meanLogEditTime,<br>medianLogEditTime, medianLogContinuingBackspaceTime,<br>stdLogContinuingBackspaceTime                                                                                      | Sustained revision     |

Table 10: Dominant features and interpreted factor themes for School 1, Grade 8, ELA writing.

Table 10 continued

| Factor | Dominant features                                                                                                                                    | Theme                    |
|--------|------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------|
| F4     | stdLogClauseBreakInsertTime,<br>stdLogBetweenWordInsertTime,<br>stdLogInWordInsertTime,<br>stdLogStartingBackspaceTime,<br>stdLogWordStartInsertTime | Edit variability         |
| F5     | medianLogMinorPunctInsertTime                                                                                                                        | Minor-punctuation rhythm |
| F6     | meanLogBeforeWordInsertTime,<br>medianLogBeforeWordInsertTime                                                                                        | Pre-word planning        |

| Factor | Dominant features                                                                                                                                                                                                                                                                                                                 | Theme                                |
|--------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------|
| F1     | meanLogInWordInsertTime,<br>medianLogInWordInsertTime,<br>meanLogBetweenWordInsertTime,<br>medianLogBetweenWordInsertTime,<br>meanLogStartingBackspaceTime,<br>medianLogStartingBackspaceTime,<br>meanLogWordStartInsertTime,<br>medianLogWordStartInsertTime,<br>meanLogClauseBreakInsertTime,<br>medianLogClauseBreakInsertTime | Quick revision / clause-level rhythm |
| F2     | overallActiveWritingTime,<br>overallNavigationTime,<br>overallReadingTime,<br>overallMenuTime,<br>nBetweenWordInserts,<br>varLogEditTime,<br>varLogStartingBackspaceTime,<br>varLogContinuingBackspaceTime                                                                                                                        | Global engagement time               |
| F3     | meanLogContinuingBackspaceTime,<br>meanLogEditTime,<br>medianLogEditTime, medianLogContinuingBackspaceTime                                                                                                                                                                                                                        | Sustained revision                   |
| F4     | stdLogInWordInsertTime, stdLogBetweenWordInsertTime,<br>stdLogWordStartInsertTime,<br>stdLogBeforeWordInsertTime,<br>stdLogMinorPunctInsertTime,<br>stdLogStartingBackspaceTime                                                                                                                                                   | Edit variability                     |
| F5     | meanLogBeforeWordInsertTime,<br>stdLogClauseBreakInsertTime                                                                                                                                                                                                                                                                       | Pre-word planning                    |
| F6     | medianLogMinorPunctInsertTime,<br>meanLogMinorPunctInsertTime                                                                                                                                                                                                                                                                     | Minor-punctuation rhythm             |

Table 11: Dominant features and interpreted factor themes for School 2, Grade 8, ELA writing.

### A.3 RQ3: Gender-Linked Feature Differences

The following tables report feature-level gender differences that remained significant after FDR correction. Negative Cliff's  $\delta$  denotes higher values for female students, whereas positive values denote higher values for male students. No significant differences were observed for School 1 Grade 8 or for School 2 Grade 7 in the ELA writing context.

| Feature                       | $p_{adj}$ | Cliff's $\delta$ |
|-------------------------------|-----------|------------------|
| nBetweenWordInserts           | < .0001   | -0.5710          |
| stdLogClauseBreakInsertTime   | 0.0003    | 0.5249           |
| overallNavigationTime         | 0.0003    | -0.5164          |
| overallActiveWritingTime      | 0.0003    | -0.5062          |
| overallReadingTime            | 0.0005    | -0.4832          |
| varLogContinuingBackspaceTime | 0.0008    | -0.4670          |
| stdLogBetweenWordInsertTime   | 0.0033    | 0.4132           |
| overallMenuTime               | 0.0033    | -0.4128          |
| varLogInWordInsertTime        | 0.0098    | -0.3710          |
| varLogEditTime                | 0.0166    | -0.3390          |
| varLogBetweenWordInsertTime   | 0.0186    | -0.3394          |
| stdLogBeforeWordInsertTime    | 0.0295    | 0.3168           |
| meanLogMinorPunctInsertTime   | 0.0295    | -0.3117          |
| varLogWordStartInsertTime     | 0.0295    | -0.3096          |
| varLogStartingBackspaceTime   | 0.0443    | -0.2913          |

Table 12: Significant gender-linked feature differences for School 1, Grade 7, ELA writing (35 male, 67 female).

| Feature                       | $p_{adj}$ | Cliff's $\delta$ |
|-------------------------------|-----------|------------------|
| nBetweenWordInserts           | < .0001   | -0.5657          |
| stdLogBetweenWordInsertTime   | < .0001   | 0.5429           |
| overallNavigationTime         | 0.0009    | -0.4585          |
| stdLogClauseBreakInsertTime   | 0.0009    | 0.4514           |
| overallActiveWritingTime      | 0.0020    | -0.4214          |
| stdLogBeforeWordInsertTime    | 0.0034    | 0.4007           |
| overallMenuTime               | 0.0095    | -0.3629          |
| stdLogInWordInsertTime        | 0.0142    | 0.3443           |
| overallReadingTime            | 0.0185    | -0.3307          |
| meanLogMinorPunctInsertTime   | 0.0382    | -0.2993          |
| medianLogWordStartInsertTime  | 0.0503    | 0.2836           |
| meanLogWordStartInsertTime    | 0.0503    | 0.2807           |
| stdLogContinuingBackspaceTime | 0.0706    | 0.2629           |
| varLogStartingBackspaceTime   | 0.0964    | -0.2446          |
| stdLogEditTime                | 0.0975    | 0.2407           |
| medianLogMinorPunctInsertTime | 0.0975    | -0.2332          |
| varLogWordStartInsertTime     | 0.0975    | -0.2243          |
| varLogEditTime                | 0.0975    | -0.2021          |

Table 13: Significant gender-linked feature differences for School 1, Grade 7, Full writing dataset (40 male, 70 female).

| Feature                      | $p_{adj}$ | Cliff's $\delta$ |
|------------------------------|-----------|------------------|
| medianLogWordStartInsertTime | 0.0264    | 0.5097           |
| meanLogWordStartInsertTime   | 0.0593    | 0.4452           |

Table 14: Significant gender-linked feature differences for School 2, Grade 7, Full writing dataset (30 male, 31 female).

| Feature                        | $p_{adj}$ | Cliff's $\delta$ |
|--------------------------------|-----------|------------------|
| meanLogBeforeWordInsertTime    | 0.0591    | 0.2782           |
| nBetweenWordInserts            | 0.0591    | -0.2742          |
| medianLogBeforeWordInsertTime  | 0.0591    | 0.2699           |
| overallNavigationTime          | 0.0591    | -0.2562          |
| overallReadingTime             | 0.0591    | -0.2523          |
| overallActiveWritingTime       | 0.0591    | -0.2494          |
| varLogEditTime                 | 0.0882    | -0.2272          |
| medianLogClauseBreakInsertTime | 0.0882    | 0.2263           |

Table 15: Significant gender-linked feature differences for School 2, Grade 8, ELA writing (75 male, 74 female).

| Feature                        | $p_{adj}$ | Cliff's $\delta$ |
|--------------------------------|-----------|------------------|
| stdLogClauseBreakInsertTime    | 0.0117    | 0.3144           |
| nBetweenWordInserts            | 0.0117    | -0.3139          |
| overallNavigationTime          | 0.0117    | -0.3114          |
| stdLogBeforeWordInsertTime     | 0.0240    | 0.2844           |
| overallMenuTime                | 0.0348    | -0.2655          |
| overallActiveWritingTime       | 0.0348    | -0.2618          |
| overallReadingTime             | 0.0521    | -0.2432          |
| stdLogBetweenWordInsertTime    | 0.0521    | 0.2402           |
| medianLogClauseBreakInsertTime | 0.0827    | 0.2208           |
| medianLogBeforeWordInsertTime  | 0.0992    | 0.2106           |

Table 16: Significant gender-linked feature differences for School 2, Grade 8, Full writing dataset (79 male, 75 female).