Handwriting correlates of the internal structure of Chinese characters

Shu-Han Lin & James Myers

National Chung Cheng University

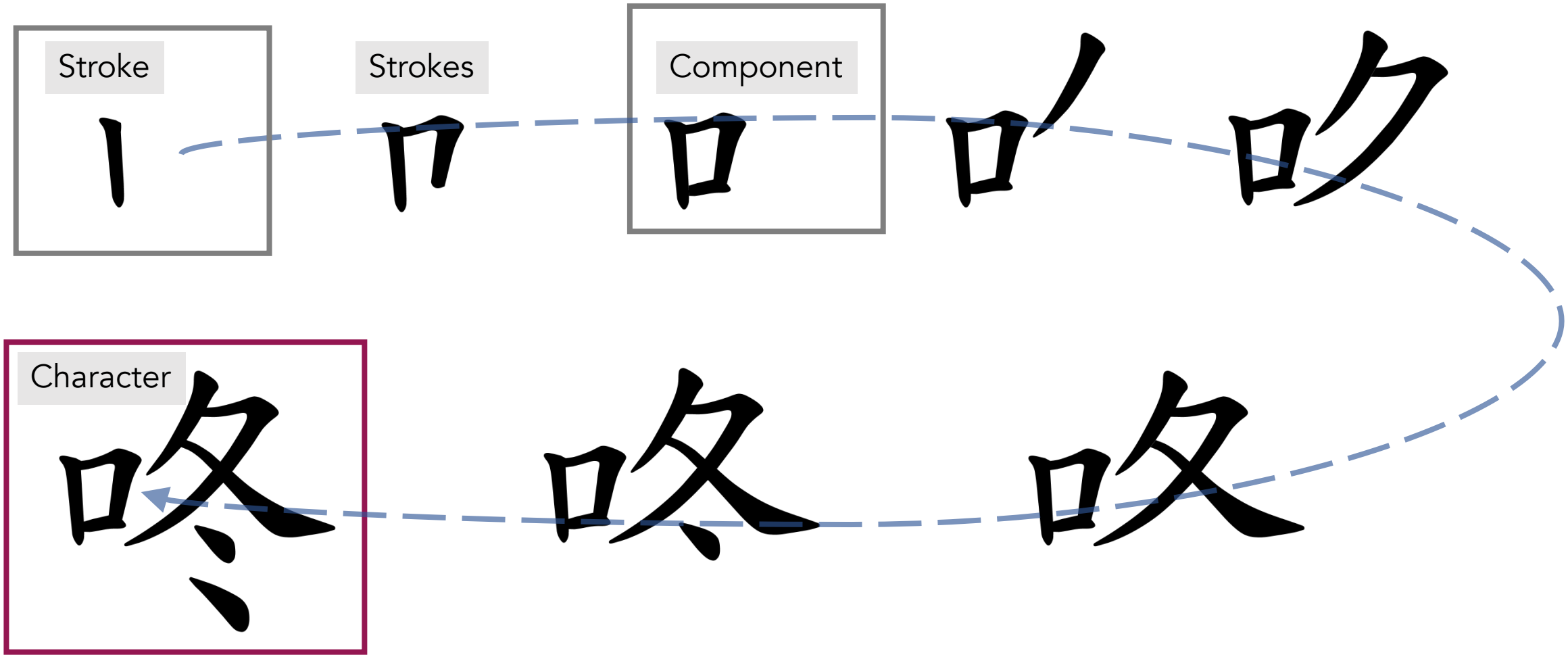
(Taiwan)

DGfS 2026: 24-27 Feb

Overview

1. Background
   - Chinese character structure
   - Inconsistencies across databases
   - “Prosody” in Chinese characters and its handwritten correlates
2. Study design
3. Step 1: Confirm final prominence in Chinese characters
4. Step 2: Apply the predictive model to ambiguous cases
5. Implications

Chinese character structure



Inconsistencies across databases

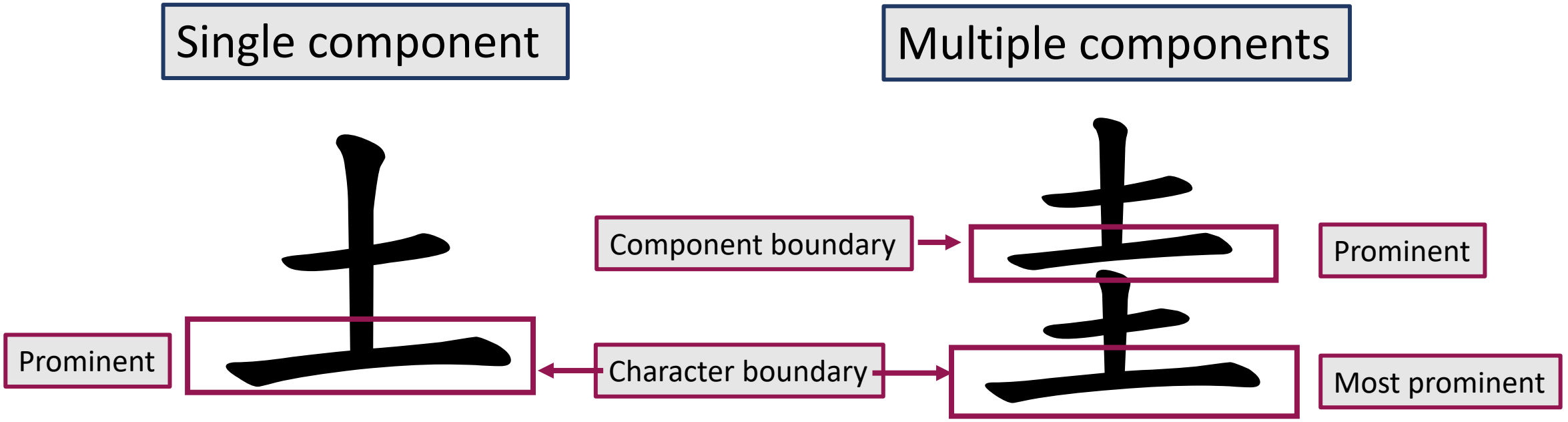
Character	ISO (Lu et al., 2002)	ASC (Chuang & Teng, 2009)	Wiki (Wikimedia Commons, 2026)	CDL (Bishop & Cook, 2007)
子	了一	子	子	子
生	丿	生	牛一	𠂉一丨一

So how do Chinese writers actually decompose characters?

Our approach: Test for handwriting correlates of character decomposition.

“Prosody” in Chinese characters

- Myers (2019)



Boundary = Final stroke → Prominent
Non-boundary = Nonfinal stroke → Non-prominent

Handwriting correlates of character "prosody"

- **Inter-stroke interval**: time between the end of one stroke and the start of the next
  - Between components > within components (Lau, 2020)
- **Speed**: length/duration of a stroke
  - Final stroke may be slowest or fastest depending on other factors (Zhang & Feng, 2017)
- **Pressure**: force against the writing surface
  - Final stroke shows greatest pressure (Kao et. al., 1986)
  - This measure is less well studied and requires further validation

Study design:
Modeling handwriting data in two steps

- Step 1-1
  - Validate handwriting correlates
- Step 1-2
  - Build predictive model with unambiguous characters

- Step 2:
  - Test model on ambiguous characters

Step 1: Characters with unambiguous structure

- Participants:
  38 native traditional Chinese writers
- Stimuli:
  150 unambiguous characters

Status	Character	ISO	ASC	Wiki	CDL
Unambiguous	月	月	月	月	月
	冬	夂冫	夂冫	夂冫	夂冫
Ambiguous	子	了一	子	子	子
	生	ノ	生	牛一	𠂉一丨一

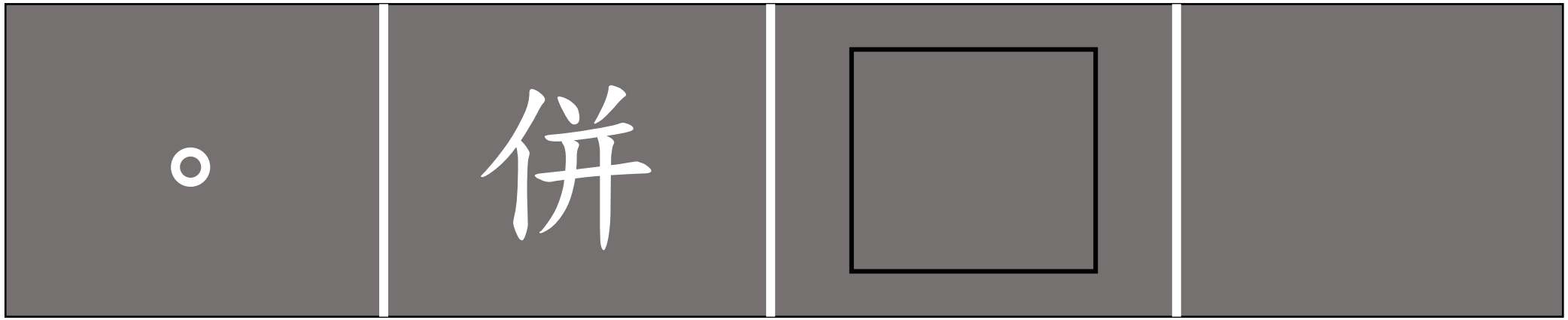
Step 2: Characters with ambiguous structure

- Participants:
  33 native traditional Chinese writers
- Stimuli:
  159 ambiguous characters

Status	Character	ISO	ASC	Wiki	CDL
Unambiguous	月	月	月	月	月
	冬	夂冫	夂冫	夂冫	夂冫
Ambiguous	子	了一	子	子	子
	生	ノ	生	牛一	𠂉一丨一

Procedure

- Wacom One® Tablet and PsychoPy (Peirce et al., 2022) recorded coordinate, time and pressure



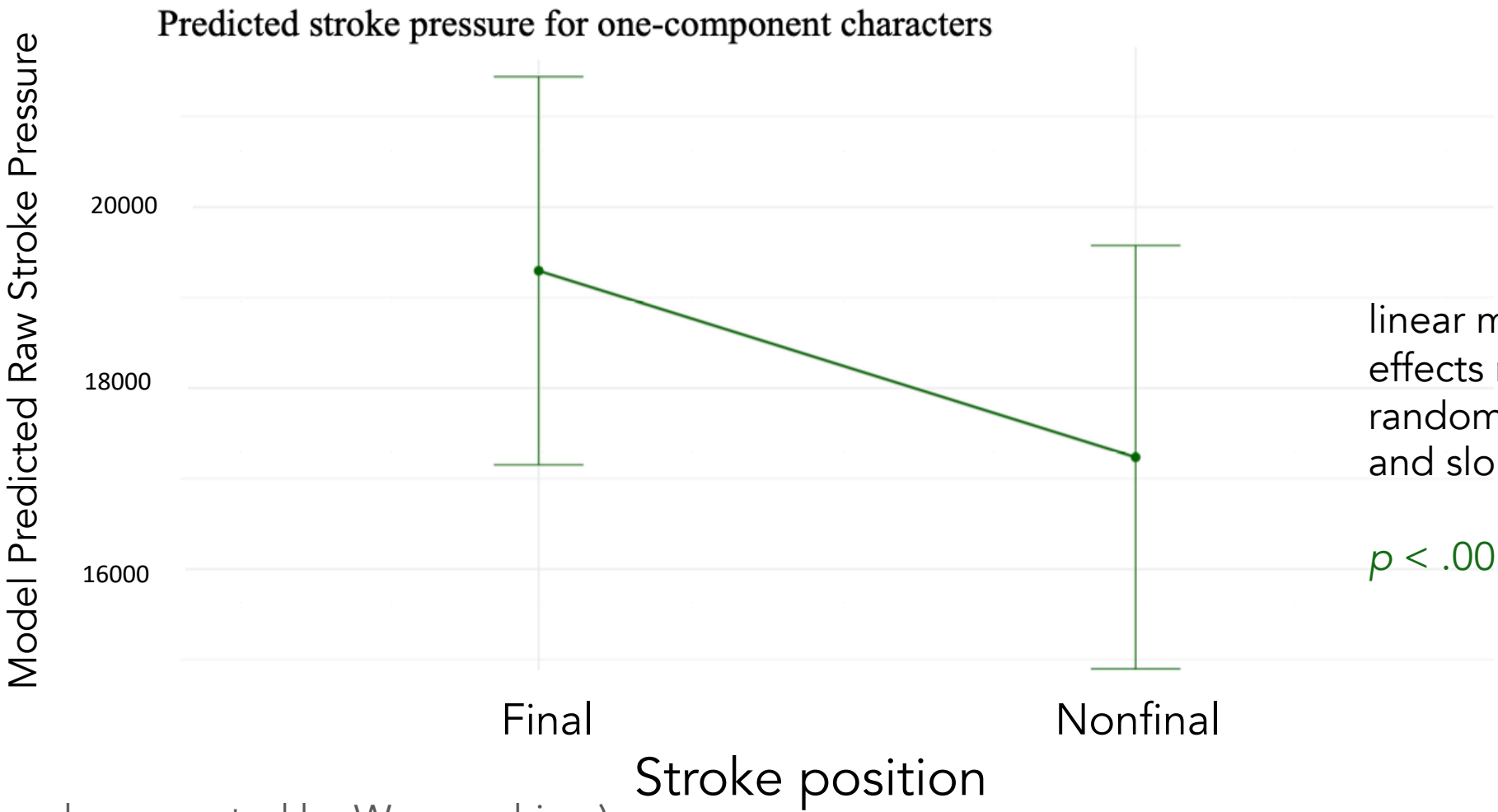
Fixation
1 sec

Stimulus
1 sec

Write stimulus
(no time limit)

Hit spacebar
to end trial

Step 1 - 1 Validate pressure as correlate: **Character**-final prominence

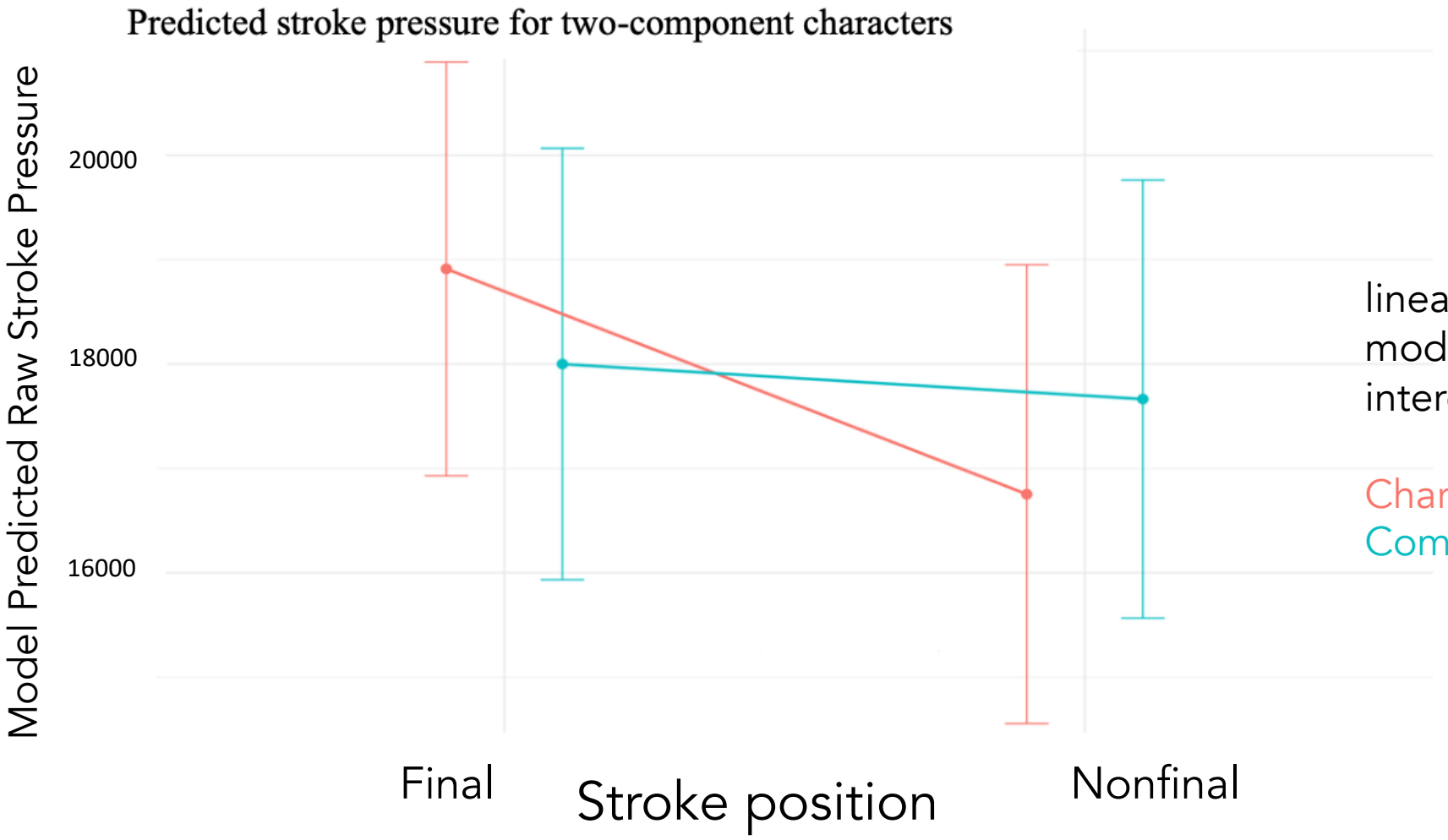


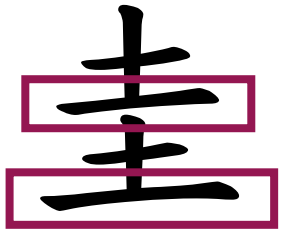
linear mixed-effects model, random intercepts and slopes

$p < .001$

(pressure scale generated by Wacom driver)

Step 1 - 1 Validate pressure as correlate: **Component**-final prominence





linear mixed-effects model, random intercepts and slopes

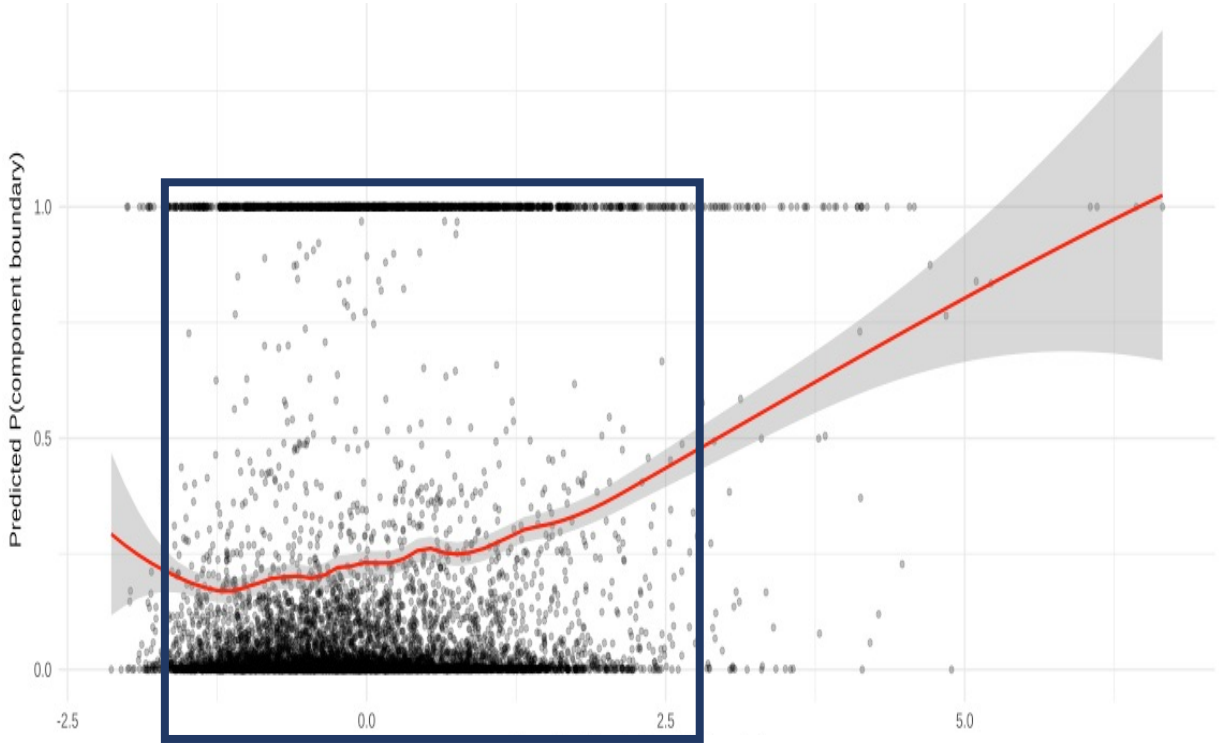
Character: $p < .001$
Component: $p = .08$

Step 1 - 2 Build predictive model for **unambiguous** characters

- Generalized additive mixed modeling (GAMM)
- Predict position: final vs. nonfinal
- Independent fixed variables (with relevant interactions):
  - **Inter-stroke interval, pressure** and **speed**
  - **Previous stroke pressure**
  - Length, width, and **stroke type**
  - Normalized stroke order (0-1) and **total stroke count**
- The best-performing model:
  - 77% sensitivity
  - 77% specificity

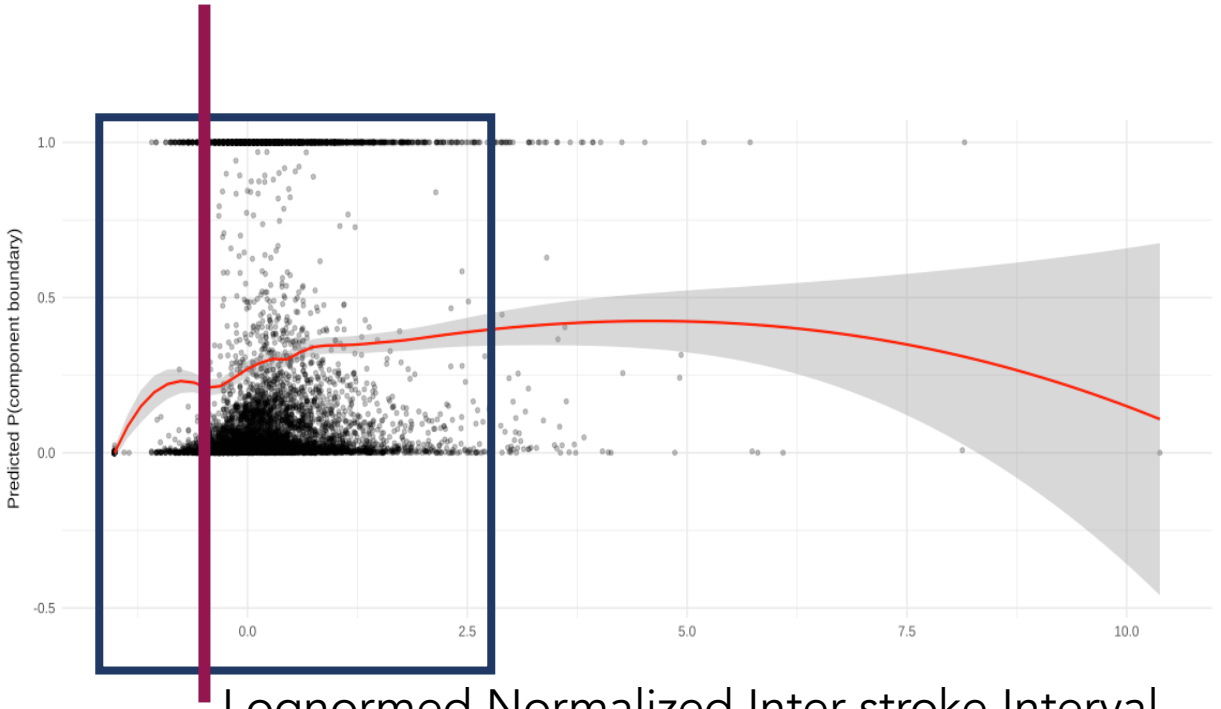
Step 1 - 2 Other correlates

Final-Fast: Contrary to "final lengthening" in speech, writing accelerates at boundaries.



Normalized Stroke Speed

Non-linear: A pause only needs to be "long enough" to signal a boundary. The probability curve flattens at longer intervals.

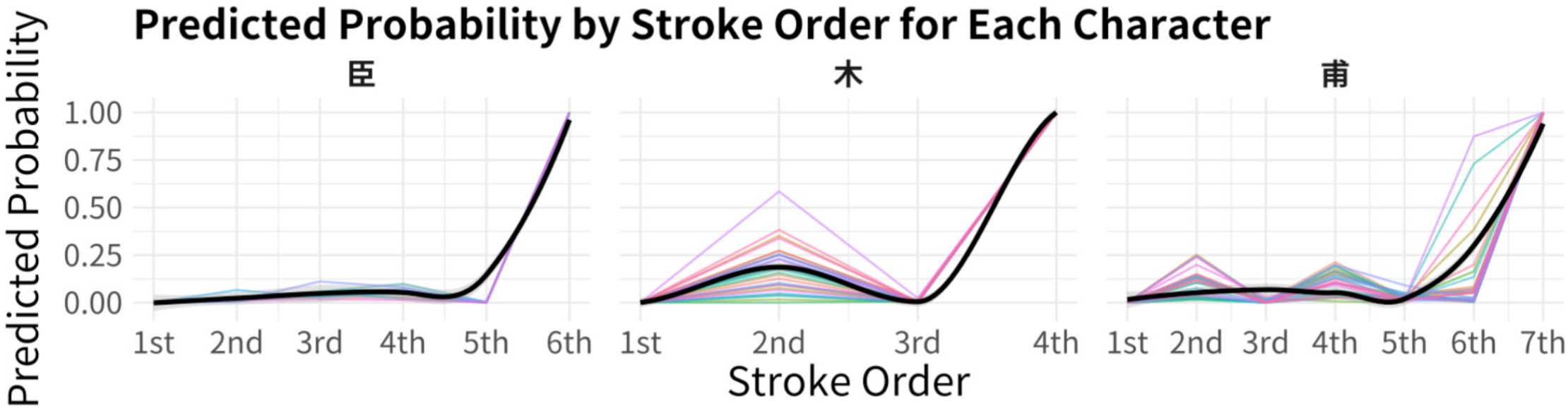


Lognormed Normalized Inter-stroke Interval

Step 2 - Apply model to **ambiguous** characters

- The results revealed three distinct categories of characters

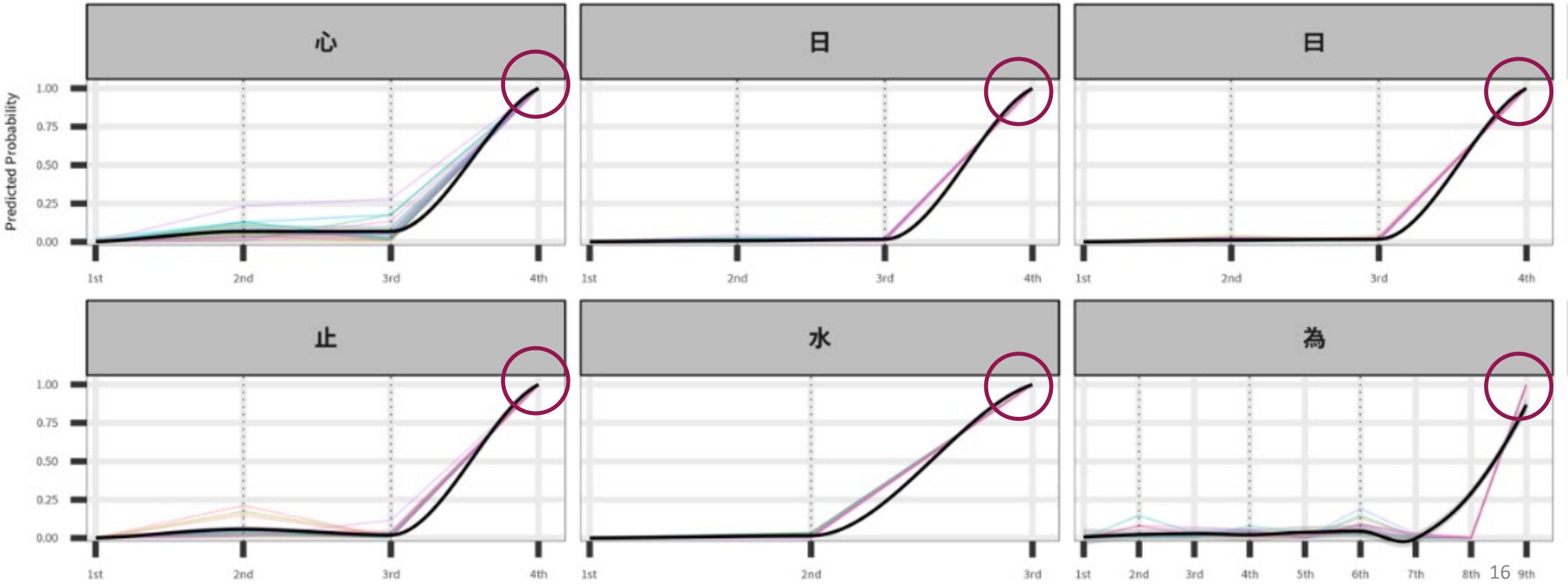
1. No internal boundary	2. Single internal boundary	3. Variation



(black = overall model prediction; coloured = individual participants)

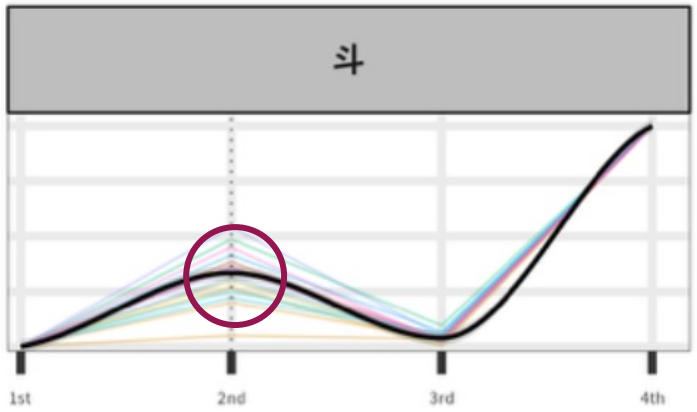
Category 1: No obvious internal boundary → one-component character

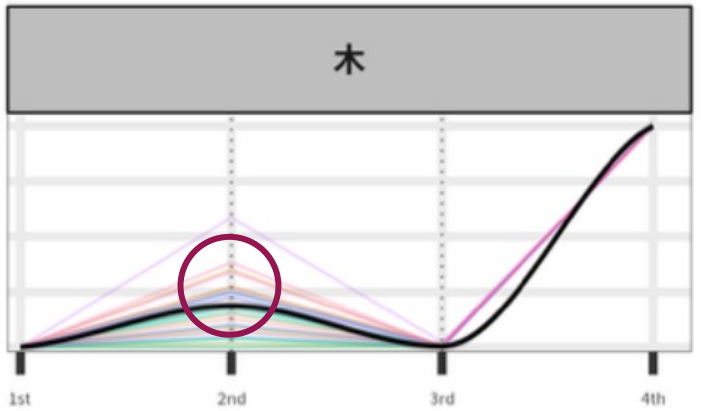
- 心、日、曰、止、水、為

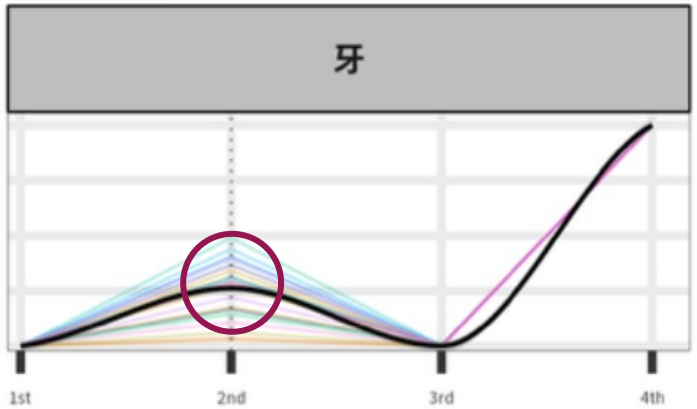


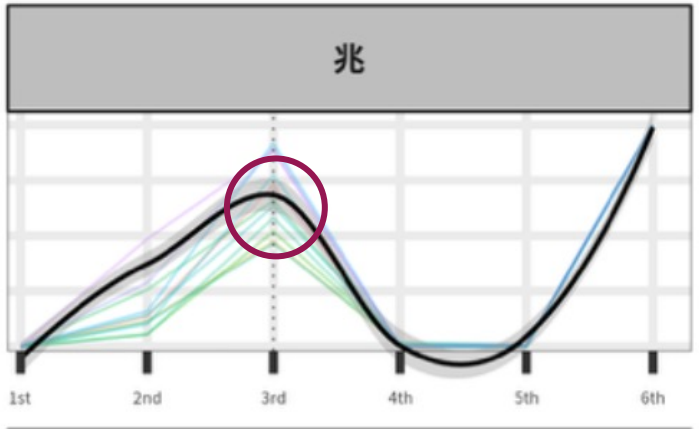
Category 2: Single internal boundary → two-component character

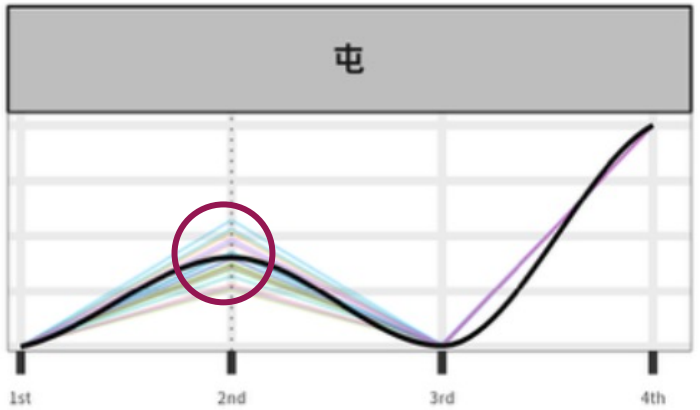
- 斗、木、牙、兆、屯、父

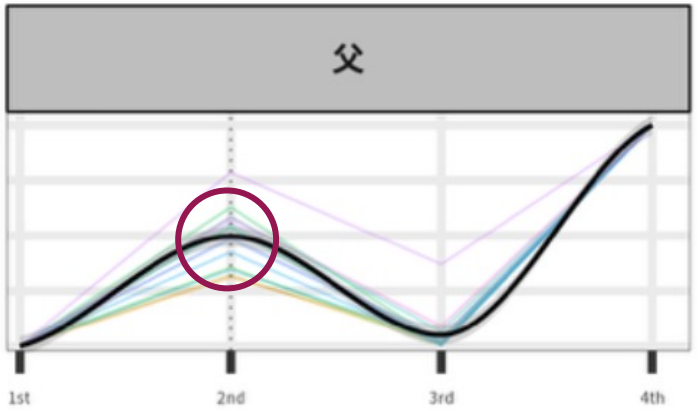




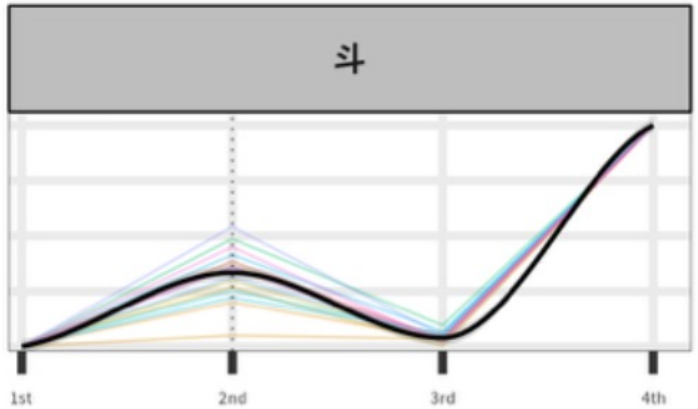


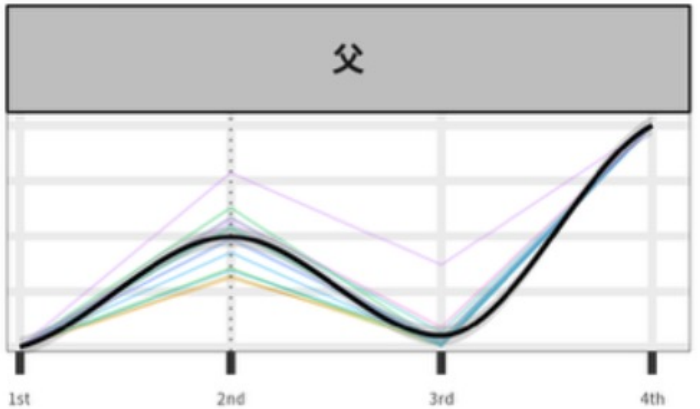


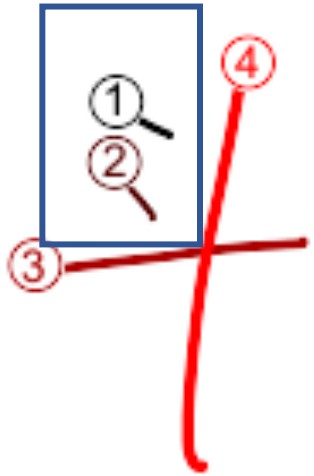


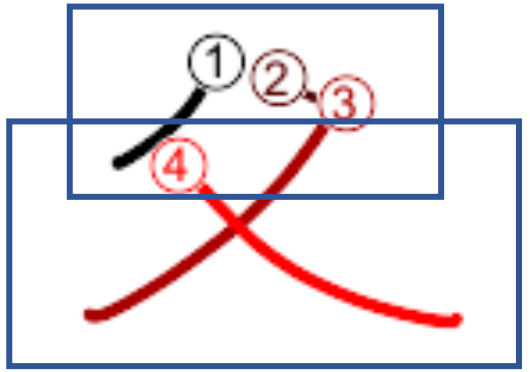


Category 2: Single internal boundary → two-component character



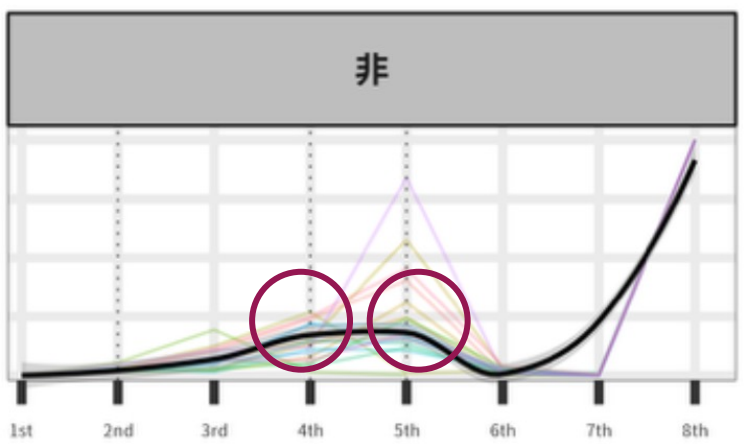


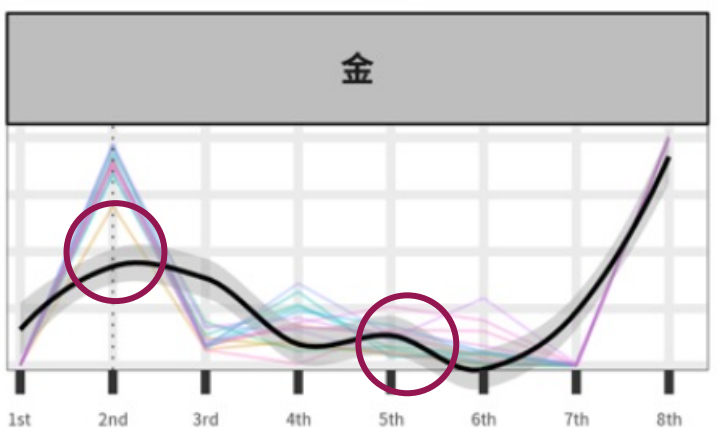


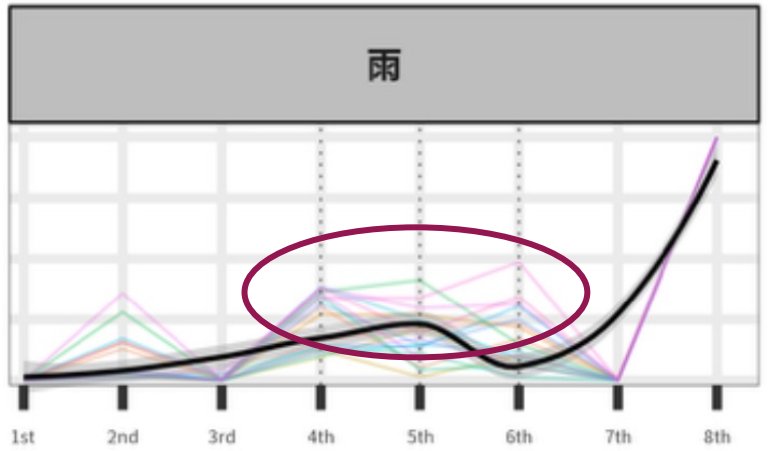


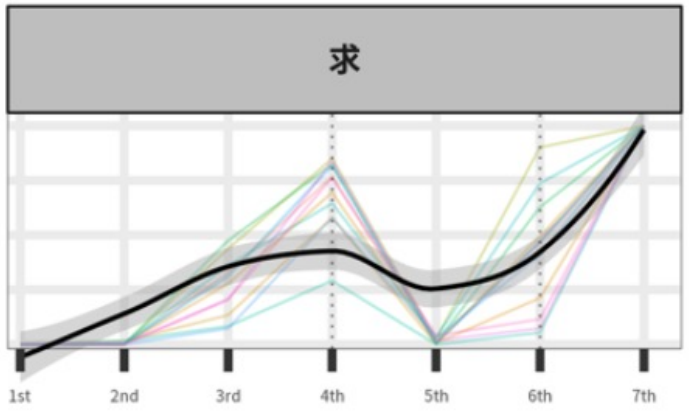
Category 3: Variation
→ different writers parsed differently

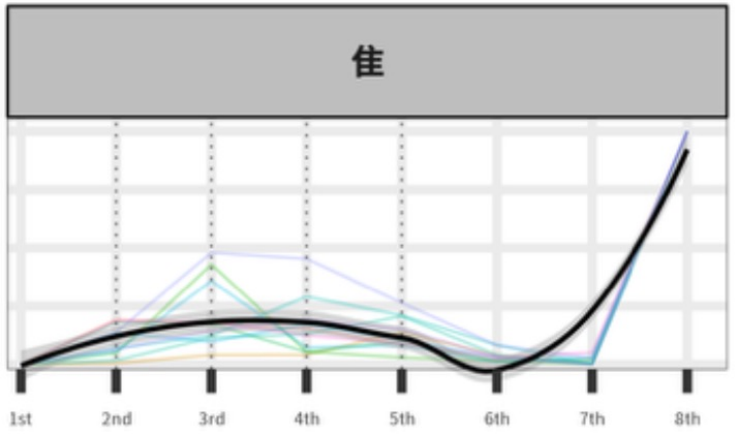
- 非、金、雨、求、准、甫

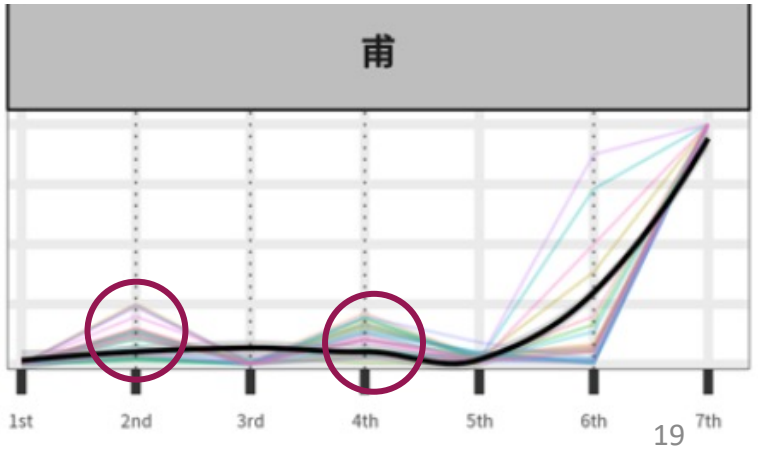




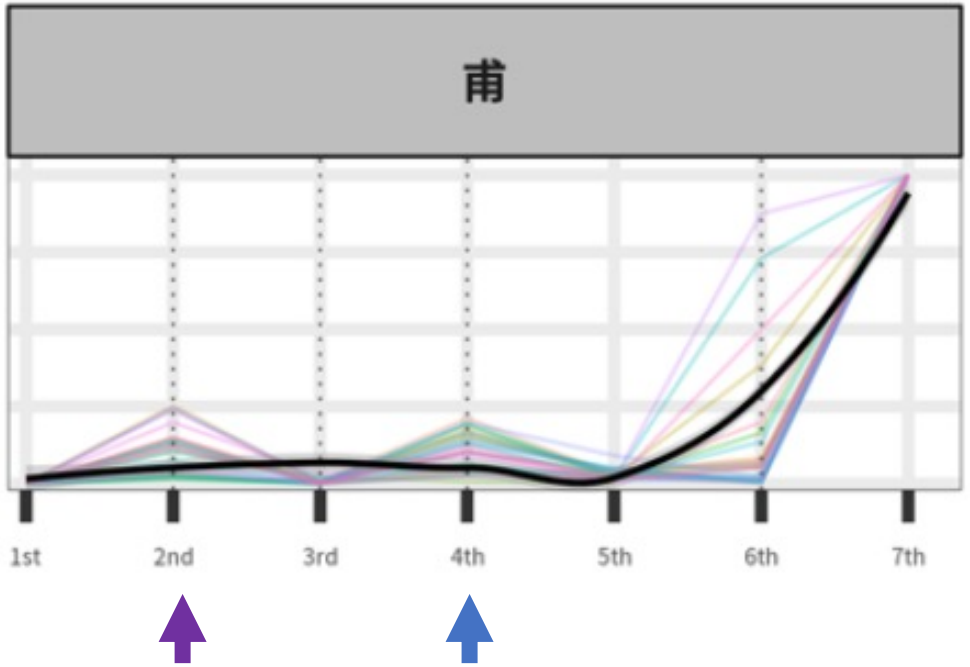


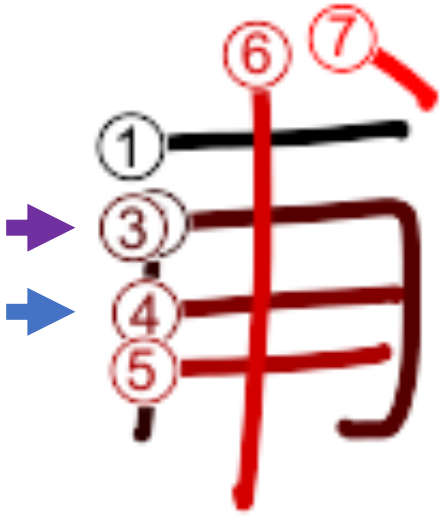






Category 3: Variation
→ different writers parsed differently





Recap

- Step 1: unambiguous characters
  - Predicting pressure by position

    Boundary = Final stroke → Strong Pressure
    Non-boundary = Nonfinal stroke → Weak Pressure
  - Predicting position by handwriting data

    The best-performing model
    77% sensitivity and 77% specificity

Step 2: ambiguous characters

- Three categories: no internal boundary, single internal boundary and variation
- The third category needs more research:
  - Variation may involve factors other than component parsing?

Implications

- Handwriting reflects character-internal structure
- This is analogous to the prosodic traces left by morphological structure in spoken words
- Our methods may be extended to other writing systems
  - Roman letters are composed of “bodies” and “codas”
  - e.g. <b> = l + ɔ (Primus, 2004)
- They may also have implications for spoken words with ambiguous morphological structure
  - helicopter = helico + pter (historical etymology)
  - heli + copter (folk etymology)

Thanks

- Taiwan's National Science and Technology Council (112-2410-H-194-030-MY3)
- 林采諺

- Code
  - Simpson, S., Nottbusch, G., & Torrance, M. (2024). OpenHandWrite [Computer software]. https://github.com/isolver/OpenHandWrite (Original work published 2015)
  - Myers, J. (2023). Nonstatistical things that linguists can do with R: A review with a case study on Chinese handwriting. In C.-J. C. Lin, A. Cherici, & B. Chen (Eds.), Proceedings of the 34th North American Conference on Chinese Linguistics (NACCL- 34). Department of East Asian Languages and Cultures, Indiana University Bloomington.

References

Bishop, T., & Cook, R. (2007). A character description language for CJK. *MultiLingual, 18*(7), 62–68.

Chuang, D.M. [莊德明], & Teng, H.Y. [鄧賢瑛] (2009). 漢字構形資料庫的研發與應用 [Research and development of Chinese characters information database and its application]. Academia Sinica, http://cdp.sinica.edu.tw/service/documents/T090904.pdf

Kao, H. S. R., M. P. Hong & L. P. Wah (1986). Handwriting pressure: Effects of task complexity, control mode and orthographic difference. *Advances in Psychology, 37*, 47–66.

Lau, D. K.-Y. (2020). Validation of constituent logographemes and radicals in Chinese characters using handwriting data. *Behavior Research Methods, 52*(1), 305–316.

Lu, Q., Chan, S. T., Li, Y., & Li, N. L. (2002). Decomposition for ISO/IEC 10646 Ideographic Characters. In *Proceedings of the 3rd Workshop on Asian Language Resources and International Standardization* (ALR 2002, COLING-02). https://aclanthology.org/W02-1209/

Myers, J. (2019). *The grammar of Chinese characters: Productive knowledge of formal patterns in an orthographic system*. London: Routledge.

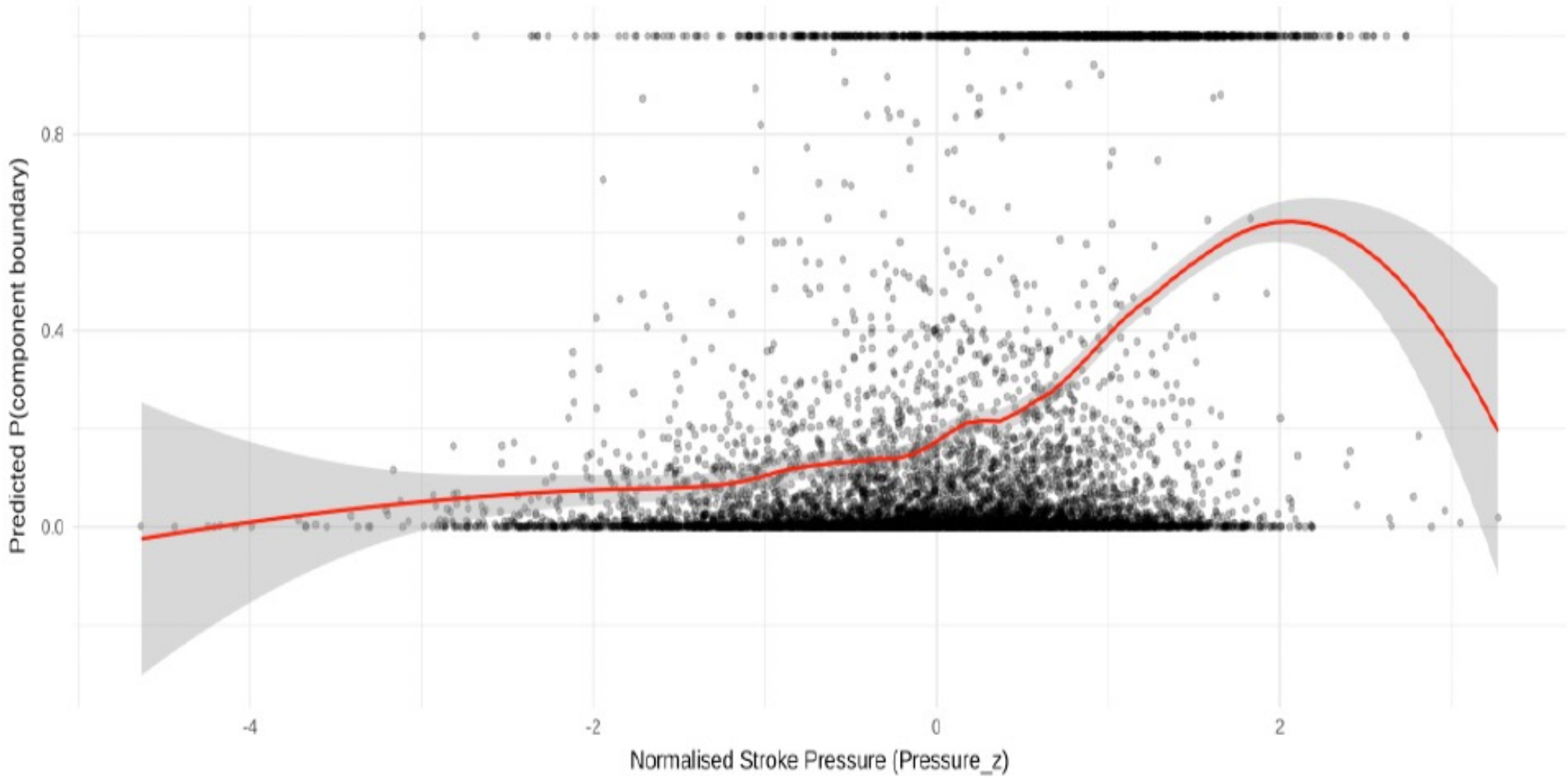
Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods, 51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

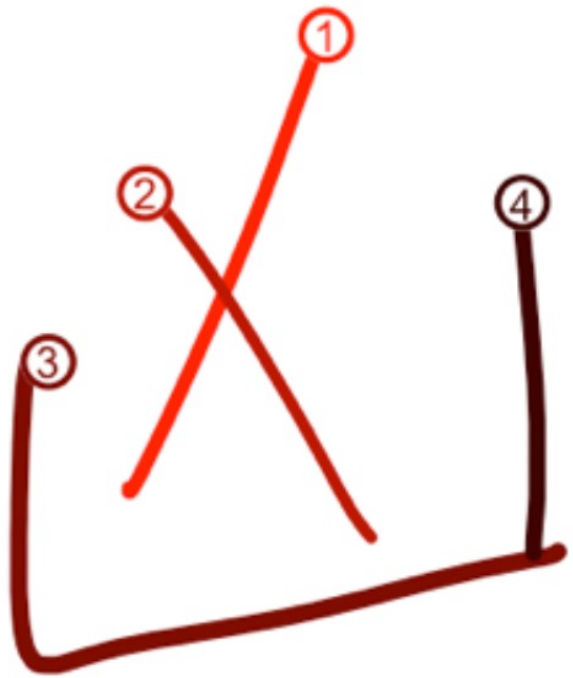
Primus, B. S. (2004). A featural analysis of the Modern Roman Alphabet. *Written Language & Literacy, 7*(2), 235-274.

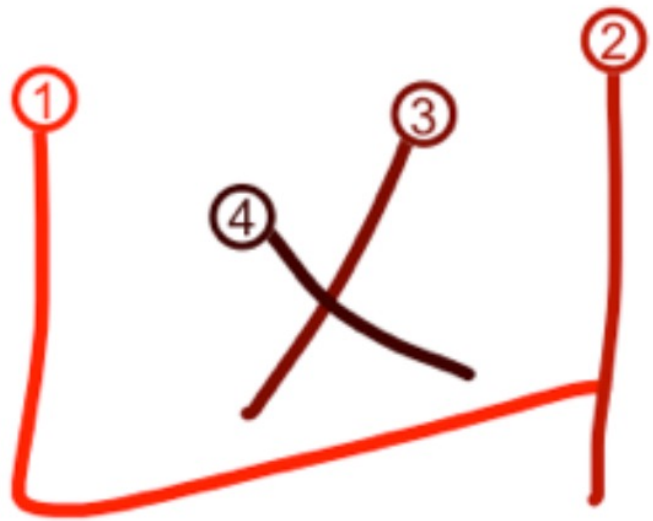
Wacom. (n.d.). Wacom help (DTC133). Retrieved August 26, 2025, from https://101.wacom.com/UserHelp/zh_TW/TOC/DTC133.html

Wikimedia Commons: Chinese characters decomposition—Wikimedia Commons. (n.d.). [Dataset]. Retrieved June 4, 2025, from https://commons.wikimedia.org/wiki/Commons:Chinese_characters_decomposition

Zhang, Q. & C. Feng (2017). The interaction between central and peripheral processing in Chinese handwritten production: evidence from the effect of lexicality and radical complexity. *Frontiers in Psychology, 8*, 334.


Predicted P(component boundary)
0.8
0.4
0.0
-4
-2
0
2
Normalised Stroke Pressure (Pressure_z)


1
2
3
4


1
2
3
4

1. Visually incorrect characters from the target character

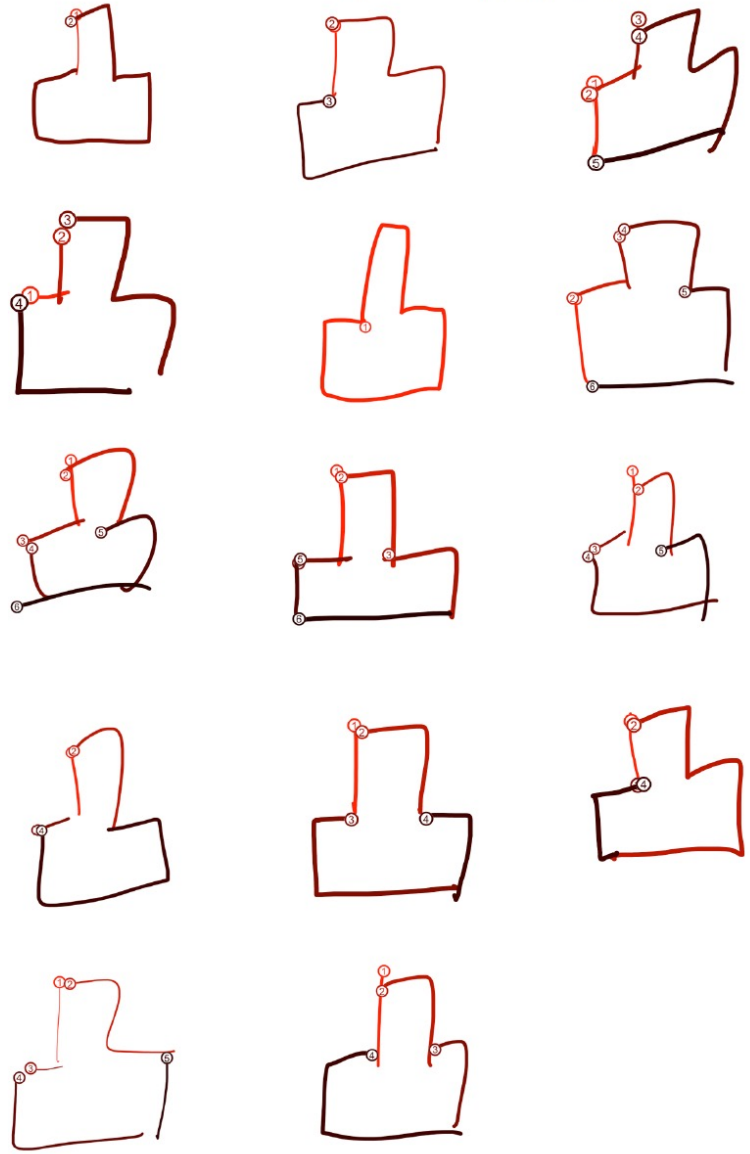
2. Wrong stroke counts (adding or missing strokes)

3. Unusual stroke orders from the majority writing patterns of participants

5. Technical Issues

Some responses were grouped together and some were separated into two.

4. Characters with fewer than three participants agreeing on the same writing order were also removed due to insufficient generalizability (e.g., 罒, 凸, 卍).



Stroke type raw pressure and frequency

Stroke Type	Stroke Type Raw Pressure	Frequency
ㄣ	23385.78	37
Z	23302.93	14
ㄟ	22407.64	42
ㄩ	21955.19	186
ㄋ	21951.26	89
ㄧ	21735.34	38
ㄣ	21536.93	19
ㄋ	21517.12	48
ㄱ	21460	111
ㄥ	21271.52	19
ㄎ	21115.27	36
ㄈ	21075.64	28
ㄦ	20665.32	22
ㄱ	20659.19	196
ㄩ	20306.45	93
ㄟ	19866.61	434
ㄱ	19321.45	604
ㄑ	19214.06	69
ㄱ	18954.54	235
ㄏ	18787.49	46
ㄙ	18691.02	28
ㄧ	18632.15	91
ㄏ	18550.64	74
ㄩ	18486.67	38
ㄱ	18197.14	737
ㄧ	18094.38	1767
—	17904.26	3471
ㄟ	17229.66	834
ㄟ	16887.31	944

One-component unambiguous characters in stimuli list for experiment 1

Character	Stroke Count	Character	Stroke Count	Character	Stroke Count
了	1	己	3	瓦	5
乙	1	工	3	母	5
一	1	川	3	乎	5
又	2	尢	3	凹	5
卜	2	女	3	凸	5
十	2	大	3	乍	5
匕	2	夕	3	世	5
力	2	士	3	承	6
刀	2	土	3	年	6
几	2	久	3	卍	6
八	2	丈	3	州	6
入	2	王	4	垂	9
人	2	片	4	才	3
二	2	火	4	弓	3
九	2	气	4	已	3
之	2	毛	4	丑	4
乃	2	毋	4	丐	4
七	2	手	4	臼	5
瓜	3	廿	4	斤	3
氏	3	井	4	五	4

Two-component unambiguous characters in stimuli list for experiment 1

Character	Stroke Count	Character	Stroke Count	Character	Stroke Count	Character	Stroke Count
詠	11	凶	4	舌	6	俚	9
郵	11	仟	5	色	6	咪	9
棵	12	冬	5	邪	6	套	9
筆	12	占	5	余	7	柚	9
策	12	叭	5	即	7	狠	9
肆	12	右	5	否	7	秋	9
補	12	外	5	吸	7	臭	9
跛	12	穴	5	坡	7	負	9
寓	12	近	5	披	7	軌	9
酥	12	件	6	攻	7	酊	9
瑪	13	先	6	斧	7	借	10
鉗	13	兵	6	杆	7	唁	10
漢	14	吋	6	男	7	恙	10
輔	14	如	6	疚	7	斬	10
燻	18	改	6	祀	7	書	10
朋	8	哄	8	祇	7	缺	10
杷	8	泄	8	阻	7	配	10
枉	8	玥	8	咀	8	針	10
迺	8	岩	8	咱	8	專	11
旻	8	佛	8	坤	8	帳	11
明	8	旺	8	妾	8	筆	11
		祂	8	委	8	裘	11

Character	Stroke Count	Character	Stroke Count	Character	Stroke Count	Character	Stroke Count
上	3	飛	7	父	4	田	5
不	4	山	3	牙	4	皮	5
丙	4	巨	5	牛	4	皿	5
且	5	巳	3	甘	5	目	5
丘	4	巴	4	甫	7	矛	5
亞	8	心	4	革	9	石	5
亦	6	文	4	韋	9	立	5
兆	6	斗	4	韭	9	羊	6
免	7	方	4	風	9	而	6
其	8	日	5	長	8	耳	6
典	8	曰	4	門	8	臣	6
函	6	曲	6	阜	7	金	8
勿	4	曳	6	隹	8	艮	6
升	4	更	7	雨	8	言	7
半	5	月	4	非	8	豆	7
卑	7	木	4	面	9	豕	6
口	3	東	9	辰	7	赤	7
寸	3	欠	4	爪	3	囧	8
小	3	止	4	尤	4	鹵	6
屯	4	民	5	尸	3	求	7
		水	3	身	6	為	9

Character	Stroke Count	Character	Stroke Count	Character	Stroke Count
的	8	恰	9	洄	9
生	5	柱	9	苓	8
至	6	盈	9	乩	6
百	6	范	8	羌	7
司	5	拋	8	怵	8
命	8	茄	8	苞	8
投	7	拾	9	侏	8
修	9	肢	8	忐	7
居	8	削	9	忑	7
苦	8	弦	8	峒	9
夜	8	亭	9	甬	7
草	9	亨	7	冽	8
批	7	刮	8	勺	3
幸	8	歪	9	皈	9
屈	8	恍	9	拄	8
券	8	咽	9	迤	8
併	8	俑	7	叟	9
狗	8	矣	7	卞	4
姊	7	沃	7	砭	9
玲	9	邸	7	伉	6
拒	7	陡	9	洽	9
柯	9	狩	9	羽	6
姿	9	郡	9	卯	5
泛	7	苟	8	佼	8
枝	8	伽	7	剃	9
茲	9				