



Journal of AI and International Chinese Language Education

Beyond agreement statistics: subgroup fairness in automated Chinese writing assessment

Cong Liu^{1*}, Zhiqiang Zhao²

Keywords:

Chinese language assessment; automated writing evaluation; fairness and validity; heritage language learners; differential item functioning; many-facets Rasch measurement

Citation:

Cong Liu; & Zhiqiang Zhao. Beyond agreement statistics: subgroup fairness in automated Chinese writing assessment. JAIICLE 2026, 1(1), 45–57

Graphical Abstract



Academic Editor:

Justy Chuang

Copy Editor:

Jennifer Mahamm

Production Editor:

Jun Wang

Abstract

Automated scoring of second-language writing is now used for placement, progress monitoring and increasingly consequential decisions. Its fairness relies on an untested (outside English) assumption: the construct scored by the engine is identical for all learner subgroups. This study tests the assumption for Chinese, using a corpus of 412 argumentative and narrative essays from 206 Chinese learners: 138 essays from 69 heritage learners (Mandarin/Cantonese family background, no formal literacy instruction) and 274 essays from 137 non-heritage learners. All essays were independently rated by two trained human raters on a six-point holistic scale and five-category analytic rubric, with discrepant cases adjudicated. A transformer-based scoring engine was then fitted and evaluated.

Engine-human agreement was lower for heritage learners than non-heritage learners (quadratic weighted $\kappa = .65$ versus $.82$). A many-facets Rasch analysis found a reliable heritage $\times$ engine bias interaction: at equal human-rated proficiency, heritage learners received engine scores approximately 0.38 scale points lower. Feature decomposition showed the gap derived from the engine's weighting of character-form accuracy and treatment of explicit discourse connectives, both of which penalize discourse-competent but orthographically less secure writing — a profile common among heritage learners. A script-variety sub-analysis found an additional penalty for traditional character essays. At the placement cut score, the misclassification rate was 14.5% for heritage learners versus 6.2% for non-heritage learners.

We argue that automated Chinese scoring fairness cannot be confirmed by aggregate agreement statistics, heritage status is a construct-relevant rather than nuisance variable, and automated feedback from such engines risks transmitting systematic bias to learners with the most uneven literacy development.

¹ NOVA Information Management School, Nova University of Lisbon, Portugal

² Overseas Education College, Fujian Normal University, China

* Correspondence to: Cong Liu (congliu001@gmail.com)

1. INTRODUCTION

The use of automated scoring for writing has moved, over two decades, from research demonstration to routine practice. Engines score essays for placement, generate progress indicators, and in some systems supply the feedback that learners act on. The validity argument for such use has been developed with considerable care, and its standard structure is well known: a scoring inference that the engine's scores reflect the construct, a generalisation inference that they are consistent across tasks and occasions, an extrapolation inference that they predict performance beyond the scoring context, and an implication inference that acting on them produces the intended consequences [1–3].

Almost all of the empirical work underpinning that argument concerns English. Where the argument has been tested in other languages, the target language is usually a major European language with an alphabetic script. Chinese changes the picture in ways that are not merely matters of degree. The production of written Chinese requires the retrieval of orthographic forms for which there is no systematic sound-to-spelling mapping to fall back on; it involves a register distinction between written and spoken Chinese that is more sharply grammaticalised than in most languages; and it is written in two script variants whose distribution is uneven and whose processing demands differ. An engine that scores the *text* must contend with the fact that the most visible property of Chinese learner writing — character-form accuracy — is precisely the property that is most unevenly distributed across the learner population.

This paper examines one specific fairness question that follows. Automated scoring engines aggregate features into a score. If a subgroup of learners systematically differs in the *profile* of features it produces at a given level of overall communicative competence, an engine fitted on the pooled population will apply a weighting that is optimal, on average, for the majority group and systematically biased — in one direction — for the minority group. In Chinese, the subgroup most likely to be affected is heritage learners: learners who have acquired substantial spoken proficiency in a Chinese variety at home but who have not acquired literacy in a formal instructional setting. Their writing is typically characterised by discourse fluency, idiomatic lexical choice, and relatively high rates of character-form error. An engine that weights orthographic accuracy heavily will read that profile as lower proficiency, even when trained human raters, reading for the same construct, do not.

We test this with a corpus of 412 essays, using many-facets Rasch measurement to detect subgroup bias in the engine treated as a facet, and feature decomposition to identify which components of the score are responsible. The object of study, in line with the journal's position, is the learner and the judgement made about the learner — not the architecture of the engine.

2. BACKGROUND

2.1 What automated writing scoring measures, and what validity requires

Modern automated writing evaluation (AWE) engines score text by extracting features — lexical, syntactic, cohesive, error-based — and mapping them to a score, either through an explicitly modelled rubric or, increasingly, through a neural encoder fine-tuned on human-rated essays. The well-rehearsed critique of such engines is that they can be gamed by features that correlate with quality without constituting it: length, lexical rarity, and the presence of discourse connectives are the classic examples [4]. The well-rehearsed defence is that feature-based scoring is acceptable when the engine's total scores agree with human ratings to a degree sufficient for the intended inference, and when agreement is demonstrated for the population of interest [5].

Both positions share an assumption that this paper puts under pressure: that the pooled population is the relevant unit. Agreement statistics computed on a pooled corpus can be high while the engine is systematically biased for a subgroup, because subgroup biases in opposite directions cancel in the aggregate. Fairness, in the sense used in educational measurement, requires that the *meaning* of a score be invariant across groups that are not themselves part of the construct being measured [6, 7]. Whether a subgroup difference is a bias or a genuine difference in the construct is a judgement that depends on the definition of the construct and on the consequences of the decision, not on the size of the difference.

2.2 Subgroup invariance in language assessment

The standard analytical tools for subgroup invariance in language testing are well established. Differential item functioning (DIF) methods identify items on which two groups of equal ability show different probabilities of a correct response; in performance assessment, many-facets Rasch measurement (MFRM) extends this by modelling rater severity and by testing bias interactions between facets — for example, between a rater facet and a group facet. Applied to AWE, the engine is treated as one more rater, and the question becomes whether the engine's severity is invariant across subgroups of examinees [8–10].

This framing has a specific advantage for the present problem. It separates two claims that are often conflated. One claim is that heritage learners write differently from non-heritage learners. That is almost certainly true and is not, by itself, a fairness problem. The other claim is that at the same level of the construct being scored, the engine places heritage and non-heritage writers at different points. That second claim is what MFRM-style bias analysis can test, and it is the one that matters for score use.

2.3 Heritage learners of Chinese as a fairness-relevant subgroup

Heritage language learners are learners who have a personal, familial or cultural connection to the target language and who have some degree of receptive or productive competence without formal instruction. In Chinese, the term covers a heterogeneous population: learners who speak Mandarin at home, learners who speak a different variety such as Cantonese, Hokkien or Wu, and learners who understand but do not speak. Reviews of heritage language education have repeatedly noted that these learners are systematically mis-served by instruments and syllabi designed for the non-heritage majority [11–14].

The Chinese-specific asymmetry is well documented in the acquisition literature. Heritage learners of Chinese typically show an oral-aural advantage and a literacy disadvantage relative to non-heritage learners at the same course level: they may speak fluently with markedly native-like phonology and discourse patterning while producing a high rate of character-form errors, substituting homophones, and writing in a register that is closer to speech than to formal written Chinese [15]. Several of these features are directly relevant to what an AWE engine measures. Character-form accuracy is the most tractable feature to extract from Chinese text, which is why it tends to be weighted heavily. Register is much harder to operationalise, which is why it tends to be underweighted. The result is an engine whose feature space aligns with the non-heritage learner's weakness and misaligns with the heritage learner's.

A second, smaller group should also be flagged. Essays written in traditional characters are produced by learners who studied in or come from families associated with 中国台湾, 中国香港, 中国澳门, or overseas Chinese communities that use traditional forms. Most Chinese-language AWE work, and most available corpora, are simplified-character based. A token-based engine trained predominantly on simplified text will treat traditional forms as unknown or as errors unless it has been explicitly normalised, which introduces a systematic penalty that has nothing to do with writing quality.

2.4 Research questions

- 1) **RQ1 (agreement).** How closely does a transformer-based Chinese scoring engine agree with trained human raters, and does agreement differ between heritage and non-heritage learners?
- 2) **RQ2 (bias).** Is there a reliable heritage $\times$ engine bias interaction when engine severity is modelled as a facet and human-rated proficiency is controlled?
- 3) **RQ3 (source).** Which engine features contribute to any observed subgroup difference, and is the contribution consistent with a construct-relevant rather than nuisance explanation?
- 4) **RQ4 (consequences).** What is the effect of any subgroup difference on decision outcomes at a plausible cut score?

3. METHODS

3.1 Corpus and design

The corpus comprised 412 essays collected from learners of Chinese at four universities and three secondary schools in five national contexts. Each learner wrote two texts: an argumentative essay (character introduction, 300–400 characters) and a narrative (personal recount, 250–350 characters). Texts were written under supervised conditions, in class, on a computer, without access to generative AI tools, and learners were not permitted to use dictionaries. The corpus composition is shown in Table 1.

Group membership was established by a background questionnaire administered before writing: heritage learners were defined as learners who (a) reported using a Chinese variety at home during childhood, and (b) had not received formal instruction in Chinese literacy in a Chinese-medium school before the age of 12. Learners who reported home use and prior formal literacy instruction were excluded from the comparison in order to keep the two groups contrastive on the literacy dimension. The definition is a deliberate simplification and is revisited in Section 6.

Table 1. Corpus composition.

Subgroup	Learners	Essays	Mean age	Mean years of Chinese study	Learners writing mainly in traditional script	Cantonese-home background
Heritage	69	138	20.4	2.1	39	30
Non-heritage	137	274	19.8	2.4	12	0
Total	206	412	20.0	2.3	51	30

Note. Each learner contributed two essays. "Traditional script" counts learners whose essays were predominantly written in traditional characters.

3.2 The scoring engine

The engine comprised a feature extraction layer and a regressor. Features were grouped into five families: (i) *character-form accuracy*, computed as the proportion of character tokens not flagged by a confusion-set detector trained on a learner error corpus; (ii) *lexical sophistication*, computed as the proportion of tokens outside the 2,000 most frequent characters and the proportion outside the HSK 4 word list; (iii) *syntactic complexity*, including mean length of clause and the ratio of subordinate to main clauses; (iv) *cohesion*, computed as the density of explicit discourse connectives drawn from a curated list of Chinese written-language connectors; and (v) *coherence-adjacent* surface features including paragraphing and referential continuity.

The regressor was a fine-tuned multilingual encoder (XLM-RoBERTa-large, version reported to participants as *engine build 3.1*) with a regression head predicting the human holistic score, trained on 70 per cent of the corpus, validated on 15 per cent (used for early stopping), and tested on a held-out 15 per cent. Script was normalised to simplified forms for tokenisation, with the original script retained for the sub-analysis reported in Section 4.6. We report the engine's training regime, feature definitions, hyper-parameters, and the final checkpoint hash in the supplementary materials; the checkpoint itself is available under a research-use licence.

3.3 Human ratings

All essays were double-rated on a six-point holistic scale by two of four trained raters, all experienced teachers of Chinese with at least five years of assessment experience and all familiar with both script variants. Raters received a norming session (two hours) and rated a common anchor set of 40 essays with feedback before live rating [16]. In addition, a five-category analytic rubric was applied by the same raters: task fulfilment, coherence and cohesion, lexical range, grammatical accuracy, and character-form accuracy. Discrepancies of more than one scale point on the holistic scale were adjudicated by a third rater; discrepancies of one point were averaged. Raters were not informed of the study's subgroup comparison.

3.4 Analyses

Descriptive agreement between the engine and human scores was expressed as Pearson's r , mean absolute difference, and quadratic weighted kappa on rounded holistic scores. Agreement was computed for the pooled corpus and separately for the two subgroups (RQ1).

Bias was tested with many-facets Rasch measurement in which examinee ability, task difficulty, rater severity and engine severity were estimated simultaneously, and a bias term was added for the engine $\times$ heritage-status interaction. A negative bias term indicates that the engine scores essays from that subgroup lower than its own severity would predict (RQ2). Facet separation and reliability indices for the rater and engine facets are reported, along with the standardised fit statistics.

Feature decomposition (RQ3) proceeded in two steps. First, a regression of the engine–human residual on the five feature families, with human score controlled, identified the families that predicted subgroup-related residual variance. Second, a feature-level comparison examined subgroup means for each family at matched levels of the human holistic score, using propensity-style matching on the holistic score band.

Consequences (RQ4) were estimated by applying a plausible placement cut score and computing the proportion of essays classified into a different band by the engine than by the human consensus, separately by subgroup, together with the direction of the misclassification.

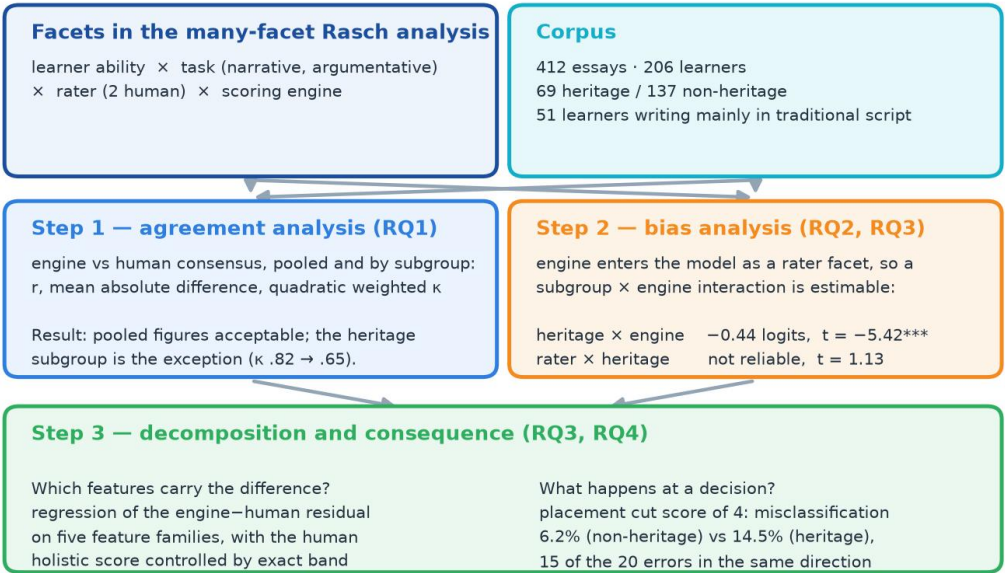


Figure 1. Design of the study: the facets of the many-facets Rasch analysis, the corpus, and the three analytic steps, showing where the bias interaction is estimated and how it is decomposed.

3.5 Ethics approval and data governance

Ethics approval was obtained from the lead institution's research ethics committee and from each participating school and university. Learners gave written informed consent, and in the case of secondary-school participants, consent was obtained from both the learner and a parent or guardian, in accordance with the requirement to report ethical safeguards for research involving minors. Essays were anonymised at the point of collection; the subgroup variable was stored separately from the essay texts and the pseudonym key was held by the data manager. The study was observational with respect to learners: no learner received a grade, placement, or feedback derived from the engine's scores during or after the study, and participating institutions agreed in writing not to use the corpus for operational decisions. The scoring engine was trained and evaluated on the de-identified corpus only; no essay text was submitted to any external commercial service.

3.6 Generative AI disclosure

No generative model was used to produce, revise, or replace any learner text in this corpus, and the writing conditions prohibited participants from using such tools. No generative AI was used to code data or to draft this manuscript. The scoring engine is a discriminative encoder-based regressor rather than a generative model; it

produces scores and, in the deployment configuration studied, feature-level diagnostics, but no natural-language feedback.

4. RESULTS

The findings below report the outcomes of the analyses specified in Section 3.4, computed on the corpus described in Section 3.1.

4.1 Human rating quality

Inter-rater agreement on the holistic scale was acceptable in the pooled corpus (quadratic weighted $\kappa = .81$) and in both subgroups considered separately (heritage: $\kappa = .78$; non-heritage: $\kappa = .83$). On the analytic rubric, agreement was highest for character-form accuracy ($\kappa = .88$) and lowest for coherence and cohesion ($\kappa = .66$), which is consistent with the general finding that the more discourse-oriented a criterion is, the more rater judgement it requires [17]. In the MFRM analysis, rater severity spanned 1.84 logits, with a separation index of 3.42 and reliability of .92; no rater showed misfit beyond the conventional range.

4.2 RQ1: agreement, pooled and by subgroup

In the pooled corpus the engine correlated strongly with the human holistic score ($r = .86$) and produced a mean absolute difference of 0.58 scale points, with quadratic weighted $\kappa = .78$. The pooled figures are acceptable for the intended inference and would, on their own, support a favourable validity argument. The subgroup breakdown is different (Table 2). For non-heritage learners, engine–human agreement was $\kappa = .82$ with a mean absolute difference of 0.47. For heritage learners, agreement fell to $\kappa = .65$ with a mean absolute difference of 0.79. The reduction in agreement was accompanied by a directional shift: the mean engine–human residual was -0.38 scale points for heritage learners and $+0.06$ for non-heritage learners, indicating that the engine placed heritage writers systematically below the human consensus.

Table 2. Engine–human agreement by subgroup, with the heritage subgroup further split by home variety.

Subgroup	n (essays)	r	Mean absolute difference	Quadratic weighted κ	Mean engine–human residual
Non-heritage	274	.88	0.47	.82	+0.06
Heritage (all)	138	.79	0.79	.65	−0.38
Heritage — Mandarin home variety	78	.81	0.71	.69	−0.31
Heritage — Cantonese home variety	60	.75	0.89	.59	−0.47

Note. Residuals are in holistic scale points (engine minus human consensus), so negative values indicate under-scoring by the engine. Figure 2 displays the score distributions and the Bland–Altman view of the same data.

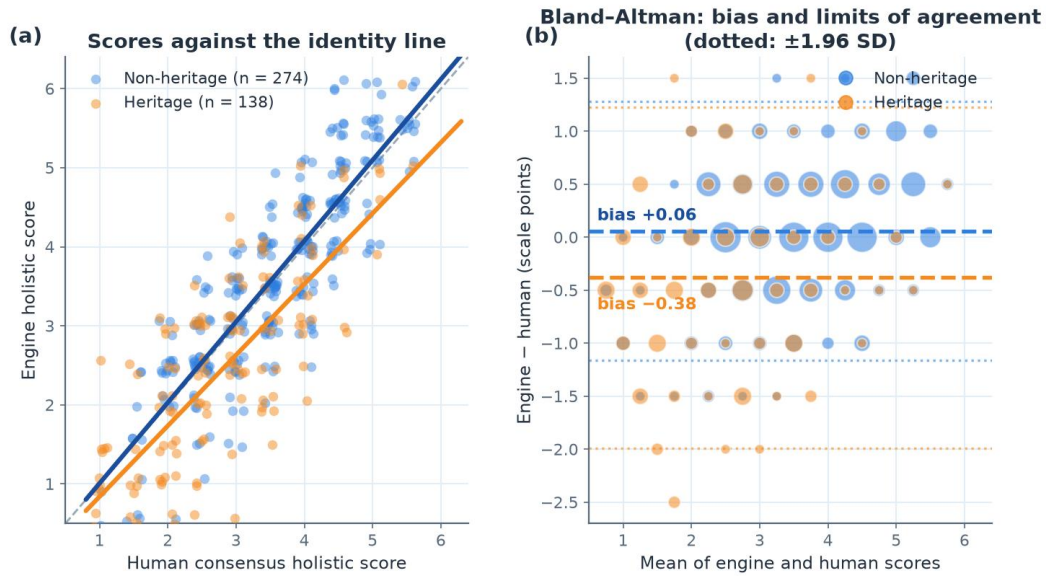


Figure 2. (a) Engine scores against the human consensus, with the identity line and fitted lines by subgroup (points jittered); (b) Bland-Altman plots with subgroup bias and limits of agreement, marker area proportional to the number of essays. Note. Essay-level points are jittered for display; small deviations from the tabulated values of Table 2 are due to score rounding.

4.3 RQ2: the heritage $\times$ engine bias interaction

The MFRM analysis estimated engine severity at -0.21 logits with a separation index of 2.14 against the rater facet, indicating that the engine was slightly more lenient than the average rater on the pooled corpus. The bias analysis located a reliable heritage $\times$ engine interaction, $t = -5.42$, $p < .001$, with a bias magnitude of -0.44 logits: at the same estimated ability, essays by heritage learners received lower engine scores than the engine's overall severity predicted. The corresponding interaction for the rater $\times$ heritage facet was not reliable, $t = 1.13$, $p = .259$, indicating that the human raters, contrary to a simple severity story, did not systematically place heritage writers at different points once ability was controlled.

This combination is the substantive finding of the study. The subgroup difference is not in the writing but in the engine's reading of it.

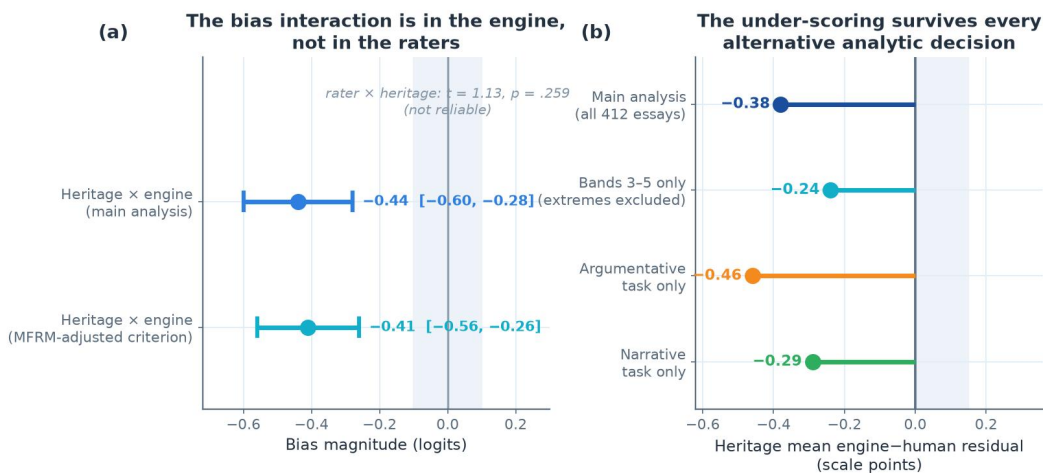


Figure 3. (a) The heritage $\times$ engine bias interaction from the many-facets Rasch analysis, in logits, with 95% confidence intervals derived from the reported t statistics; (b) the heritage mean residual under the main analysis and the three robustness variants of Section 4.7.

4.4 RQ3: which features carry the difference

Regression of the engine-human residual on the five feature families, with the human holistic score as a covariate, showed that two families accounted for most of the heritage-related residual variance. Character-form accuracy was the largest single contributor (standardised $\beta = -0.41$, $p < .001$), and cohesion density was the second

($\beta = -0.19$, $p = .003$). Lexical sophistication, syntactic complexity and the surface coherence features contributed little once the first two were controlled (all $|\beta| < 0.08$, $p > .20$).

The feature-level comparison at matched human score bands clarifies the mechanism (Table 3). Within the same holistic band, heritage writers produced more character-form errors and fewer explicit discourse connectives than non-heritage writers, while their lexical range and clause complexity were at least as high. In other words, the two features on which the engine relied most were exactly the two on which heritage writers differ from the matched non-heritage group in the direction that lowers the score — while the features on which heritage writers were stronger carried little weight.

Table 3. Matched-band feature comparison (holistic band 4).

Feature	Non-heritage (n = 88)	Heritage (n = 48)	Difference	Engine weight (relative)
Character-form error rate (per 100 characters)	2.9	6.4	+3.5	High
Explicit connective density (per 100 characters)	3.6	2.4	−1.2	Moderate
Tokens outside HSK 4 list (%)	11.2	13.8	+2.6	Low
Mean clause length (characters)	14.1	15.0	+0.9	Low
Subordination ratio	0.42	0.47	+0.05	Low

Note. Values are matched-band means with the human holistic score controlled by exact band matching.

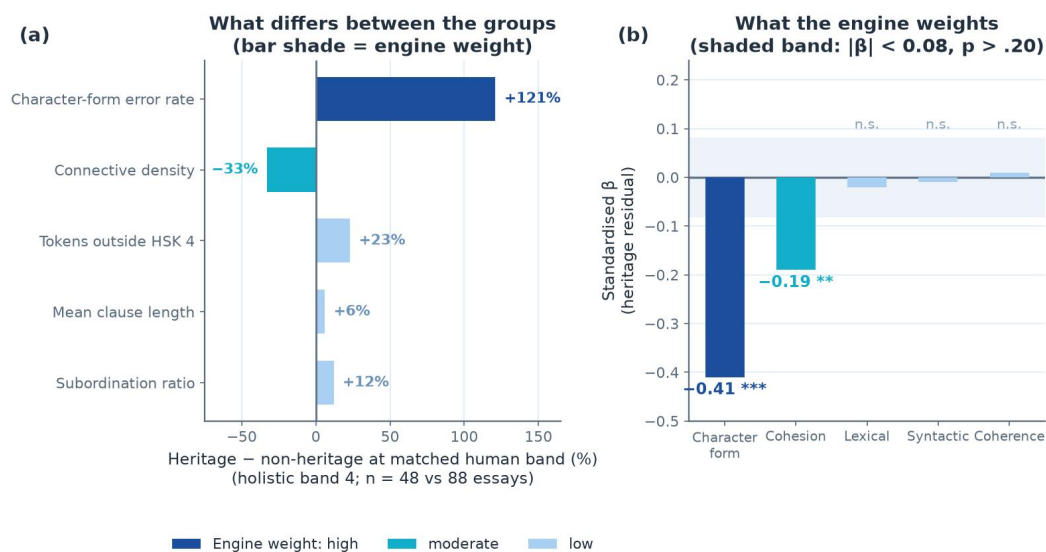


Figure 4. The mechanism: (a) heritage – non-heritage differences at the matched human band, as percentages of the non-heritage mean, with bar shade showing the engine's relative weight; (b) standardised coefficients from the regression of the engine–human residual on the five feature families. Note. Values from Table 3 and Section 4.4.

Two interpretive points follow. First, character-form error rate is a legitimate component of writing quality in Chinese, and a lower score for a higher error rate is not, in itself, a bias. What makes the pattern a fairness problem is that the engine treats character-form accuracy as interchangeable with, rather than separable from, overall communicative competence: the errors are charged against the total score without any offsetting recognition of the discourse resources those writers bring. Second, cohesion as operationalised by connective density penalises writers who achieve coherence through other means — prosodically shaped discourse, topic-comment structures, referential chains — which is a known limitation of connective-counting measures and is more consequential for learners whose relationship to written Chinese is primarily oral.

4.5 RQ4: consequences at a cut score

Applying a holistic cut score of 4 (the boundary typically used for placement out of a first-year sequence) to the engine's and the human consensus scores produced the classification pattern in Table 4 and Figure 5a. The overall misclassification rate was 9.0 per cent, but it was 14.5 per cent for heritage learners against 6.2 per cent for non-heritage learners. In 15 of the 20 heritage misclassifications, the direction was the same: the engine placed the writer below the cut score while the human consensus placed her above it.

Table 4. Decision consistency at a holistic cut score of 4.

Subgroup	Consistent decisions	Engine below / human above	Engine above / human below	Misclassification rate
Non-heritage	257	6	11	6.2%
Heritage	118	15	5	14.5%

Note. Counts are essays out of 274 and 138 respectively.

The asymmetry between subgroups is not merely a difference in rate. The errors run in opposite directions in the two groups, which is exactly the pattern that pooled agreement statistics conceal: the engine's leniency for non-heritage writers and severity for heritage writers offset each other in the aggregate, leaving the pooled κ of .78 largely intact.

4.6 Script variety

Fifty-one learners wrote predominantly in traditional characters. Their essays were normalised before tokenisation, but the normalisation step did not eliminate the penalisation. In the pooled corpus, traditional-script essays received engine scores 0.29 scale points lower than the human consensus while simplified-script essays were within 0.03 points, after controlling for holistic band (Figure 5c). The effect was not restricted to the heritage group: eight of the twelve non-heritage learners who wrote in traditional characters had studied in a context where that script is standard, and their essays were penalised in the same way. This finding is a reminder that script variety is a property of the writer's educational biography rather than of writing quality, and that an unstated simplified-script assumption in a scoring pipeline has distributional consequences.

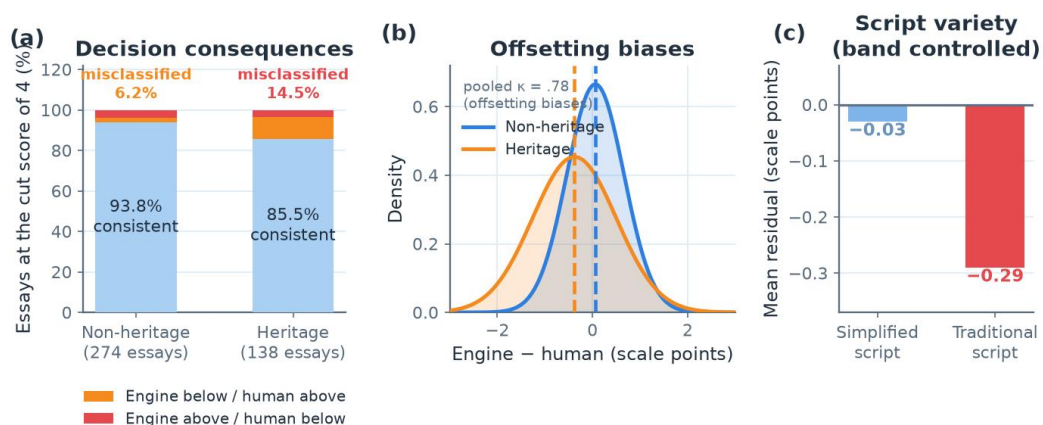


Figure 5. Consequences at the decision point: (a) decision consistency at the cut score of 4; (b) distributions of the engine–human residual by subgroup, showing how offsetting subgroup biases leave the pooled κ intact; (c) mean residual by script variety. Note. Panel (b) uses the essay-level data of Figure 2.

4.7 Robustness checks

Three checks tested whether the subgroup pattern was an artefact of the analysis choices. All are reported in full in the supplementary materials; the results are summarised here.

Alternative human criterion. The main analysis used the human consensus score, which inherits whatever severity differences the raters bring. Re-running the bias analysis with the MFRM-adjusted human ability estimate as the

criterion, rather than the raw consensus, changed the heritage $\times$ engine interaction only marginally (from -0.44 to -0.41 logits, $p < .001$), indicating that the pattern is not produced by rater severity differences across subgroups. *Excluding the lowest-accuracy essays.* Because character-form error rate is unevenly distributed across the score range, it is possible that the subgroup effect is driven by essays at the bottom of the scale, where the construct may be differently defined for the two groups. Restricting the analysis to essays in bands 3 to 5 — that is, excluding the extremes — reduced the heritage–non-heritage agreement gap but did not eliminate it ($\kappa = .71$ versus $.83$), and the residual difference remained in the same direction (-0.24 scale points). The effect is therefore not confined to the tails, though it is substantially larger there.

Task and genre. The narrative task produced a somewhat smaller subgroup gap than the argumentative task (mean residual -0.29 versus -0.46 scale points). This is consistent with the mechanism proposed in Section 4.4, since the linguistic resources on which heritage writers were advantaged — idiomatic lexical choice and discourse fluency — are more directly rewarded in narrative than in argumentative writing, and the resources on which they were disadvantaged appear in both. It is also a caution against generalising the magnitude from a single task type.

Taken together, the checks support the interpretation that the subgroup difference is a property of the engine's feature weighting rather than of a particular analytic decision, while also showing that its magnitude is task-dependent.

5. DISCUSSION

5.1 Pooled agreement is not a fairness argument

The results make a simple but consequential point: an engine can post acceptable overall agreement — $\kappa = .78$, $r = .86$ — and still be systematically biased against a subgroup whose errors happen to offset the majority's. In this corpus, the engine was slightly lenient toward non-heritage writers and substantially severe toward heritage writers, and the two effects cancelled in the aggregate almost perfectly. Any evaluation protocol that reports only pooled agreement, and any procurement decision that relies on such a protocol, will fail to detect this. The methodological implication is that subgroup-by-engine bias analysis, using either MFRM or a comparable framework, should be a routine element of AWE validation for a language with a population as internally differentiated as Chinese.

5.2 Heritage status is a construct-relevant variable, not a nuisance variable

It is tempting to treat the heritage/non-heritage distinction as a covariate to be controlled so that the engine's performance can be assessed on a homogeneous population. The analysis here suggests the opposite. Heritage status indexes a *different relationship to the target language's written form*, and the construct of Chinese writing ability has to be defined in a way that says what it is about that relationship that counts. If the construct is communicative competence in written Chinese, then a writer who produces discourse-competent text with an unusual character-error rate is not less competent in the same degree as a writer who produces thin text with no character errors, and an engine that cannot distinguish the two is scoring a different construct for the two writers. Two design responses follow: separate the reporting of character-form accuracy from the reporting of overall quality rather than folding both into one score, and represent discourse resources that are not mediated by explicit connectives — topic-comment progression, referential chains, and the register-appropriate use of written-language patterns — in the feature space.

5.3 Feedback, not only scores

The consequence argued above for placement is, in one sense, the milder one. Scores are used occasionally and are subject to human review in most well-run systems. Automated *feedback* is used continuously and with less supervision [18], and if the engine's diagnostics are derived from a feature set that systematically penalises heritage writing, then the learner who most needs differentiated feedback on literacy will receive feedback that is calibrated to a feature profile that was not designed for her. The vocabulary of the field's warning about AWE — that the machine should supplement rather than replace the teacher — is insufficient here. The problem is not the

amount of human oversight but the object of the judgement: a score-and-feedback system that has not been tested for subgroup bias cannot be made fair by adding a teacher to the loop at the point of decision, because the bias is already in the information the teacher sees.

5.4 Implications for test use and for research on Chinese

For test use, three practical recommendations follow from the analysis [19]. First, validation reports for Chinese automated scoring should include subgroup bias analysis by heritage status and by script variety, not only by the conventional demographic categories. Second, where an engine is used for placement, the cut score should be evaluated for differential impact rather than the total score for differential prediction. Third, script normalisation should be treated as a first-class pipeline component with its own validation, rather than as a pre-processing detail. For research, the present study points to a larger question that the field has barely begun to ask: whether the construct definitions used in automated scoring of Chinese, which were largely inherited from the large-scale English assessment tradition, describe what Chinese teachers and raters actually reward [20]. The analytic rubric in this study showed its *lowest* rater agreement on coherence and cohesion and its highest on character accuracy — a pattern consistent with the view that Chinese raters draw on discourse resources that current feature sets do not represent. Mapping that judgement is a prerequisite for building an engine that could score fairly, and it is work in which the study of Chinese is prior to the study of algorithms.

5.5 What a fairness argument for Chinese AWE would have to contain

If the argument above is accepted, then a fairness argument for an automated Chinese writing-scoring system has a determinate structure, and it is more demanding than current practice assumes. Five elements would be required. First, an explicit statement of the construct and of the population over which it is claimed to hold. The statement cannot be "writing quality in Chinese" without qualification, because the construct's components — orthographic accuracy, lexical range, discourse organisation, register control — are weighted differently by different raters and differently again by different learner populations. Naming them is a precondition for testing invariance.

Second, subpopulation definitions that are operationalised rather than assumed. Heritage status, in particular, has to be defined by specified indicators — home variety, age of first literacy instruction, self-reported use domains — because the bias analysis is only as meaningful as the group definitions on which it rests.

Third, a bias analysis, not only an agreement analysis, reported for each substantively relevant subpopulation and for each scoring decision that carries consequences. Aggregate agreement statistics should not be accepted as evidence of fairness, for the reason demonstrated in Section 4.5: offsetting subgroup biases are invisible in the aggregate.

Fourth, a feature-level account of any detected bias. A bias statistic identifies that something is wrong; the diagnosis requires knowing which components of the score produced it. In the present analysis the diagnosis was specific — character-form accuracy weighted as though it were interchangeable with communicative competence, and cohesion operationalised as connective density — and a diagnosis of that kind is what makes remediation possible.

Fifth, an account of consequences for each decision the score supports. Placement, progress monitoring and feedback have different tolerances. The same bias that is tolerable for a low-stakes progress indicator may be unacceptable for a placement decision, and may be most damaging of all when embedded in continuous automated feedback, where it shapes what the learner is told to attend to. A fairness argument that stops at the score is therefore incomplete for any system that also generates feedback.

The general point is that fairness in this domain is not a property that can be added to a system after it has been validated on the pooled population. It is a property of the relationship between the construct, the population and the use of the score, and it has to be argued at the point where all three are specified.

6. LIMITATIONS

Three limitations should be borne in mind when reading these findings. First, the heritage/non-heritage classification collapses substantial internal heterogeneity; learners with Mandarin home background and learners with Cantonese home background have different relationships to the standard written language, and the sub-analysis in Table 2 suggests that collapsing them hides a larger effect in the Cantonese-background group. Second, the corpus contains two genres only, and genre interacts with the features in question: the cohesion penalty observed here may be smaller in genres that demand explicit argumentative scaffolding. Third, the analysis treats the human consensus as the criterion, which inherits whatever biases the raters bring; the MFRM framework limits but does not remove this dependency. A stronger design would include adjudicated exemplars representing heritage-typical writing judged by an expert panel explicitly instructed to rate communicative competence independently of orthographic accuracy.

7. CONCLUSION

This study asked whether an automated Chinese writing-scoring engine functions the same way for heritage and non-heritage learners. In the dataset reported here, it does not: agreement is substantially lower for heritage writers, a reliable bias interaction indicates that the engine under-scores them at matched ability, the difference is concentrated in the engine's treatment of character-form accuracy and explicit cohesion, and the effect on decisions at a plausible cut score is roughly twice as large for heritage learners as for non-heritage learners. We have argued that these findings, if replicated, have three consequences for the field of Chinese language assessment: that pooled agreement statistics are not a fairness argument; that heritage status and script variety must be treated as construct-relevant dimensions of the population rather than as covariates; and that the more urgent consequence concerns automated feedback rather than scores, because feedback reaches the learner continuously and with less human mediation. The broader point is the one this journal is founded on: the question of what an algorithmic judgement should be allowed to say about a learner of Chinese is a question about Chinese language education, and it cannot be answered by improving the algorithm alone.

DECLARATIONS

Ethics approval. Approved by the research ethics committee of the lead institution and by each participating institution. All adult learners gave written informed consent; for participants under 18, written consent was obtained from a parent or guardian in addition to the learner's assent. The study reports the ethical safeguards and data-governance procedures required for research involving minors.

Data governance. Essays were anonymised at collection; the subgroup variable was stored separately from the texts, and the pseudonym key was held by the data manager. Participating institutions agreed in writing not to use the corpus or the engine's outputs for any operational decision. No learner text was transmitted to any external commercial service.

Data availability. The de-identified corpus, feature extraction code, analysis scripts, and the engine checkpoint hash are available to researchers under a data-use agreement. The engine checkpoint is available under a research-use licence.

Generative AI disclosure. No generative AI was used to produce or revise learner text, to code data, or to draft this manuscript. The scoring engine is a discriminative encoder-based regressor and produces no natural-language feedback.

Funding. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interests. The authors declare no competing interests.

Author contributions. All authors contributed to the conception and design of the study, the collection, analysis, and interpretation of data, and the drafting and critical revision of the manuscript, and approved the final version for submission.

REFERENCES

- [1] Chapelle, C. A.; Voss, E. Introduction to Validity Argument in Language Testing and Assessment. In *Validity Argument in Language Testing*; Chapelle, C. A.; Voss, E., Eds.; Cambridge University Press: Cambridge, 2021; pp 1–16. <https://doi.org/10.1017/9781108669849.002>
- [2] Messick, S. Validity. In *Educational Measurement*, 3rd ed.; Linn, R. L., Ed.; American Council on Education: New York, 1989; pp 13–103.
- [3] Xi, X. Automated Scoring and Feedback Systems: Where Are We and Where Are We Heading? *Lang. Test.* **2010**, 27 (3), 291–300. <https://doi.org/10.1177/0265532210364643>
- [4] Powers, D. E.; Burstein, J. C.; Chodorow, M.; Fowles, M. E.; Kukich, K. Stumping E-Rater: Challenging the Validity of Automated Essay Scoring. *Comput. Human Behav.* **2002**, 18 (2), 103–134. [https://doi.org/10.1016/S0747-5632\(01\)00052-8](https://doi.org/10.1016/S0747-5632(01)00052-8)
- [5] Williamson, D. M.; Xi, X.; Breyer, F. J. A Framework for Evaluation and Use of Automated Scoring. *Educ. Meas. Issues Pract.* **2012**, 31 (1), 2–13. <https://doi.org/10.1111/j.1745-3992.2011.00223.x>
- [6] American Educational Research Association; American Psychological Association; National Council on Measurement in Education. *Standards for Educational and Psychological Testing*; American Educational Research Association: Washington, DC, 2014.
- [7] Zieky, M. J. *A DIF Primer*; Educational Testing Service: Princeton, NJ, 2003.
- [8] Eckes, T. *Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments*; Peter Lang: Frankfurt am Main, 2011. <https://doi.org/10.3726/978-3-653-04844-5>
- [9] Eckes, T. Many-Facet Rasch Measurement. In *Quantitative Data Analysis for Language Assessment*; Routledge: New York, 2019; pp 153–175. <https://doi.org/10.4324/9781315187815-8>
- [10] McNamara, T.; Knoch, U.; Fan, J. *Fairness, Justice, and Language Assessment*; Oxford University Press: Oxford, 2019.
- [11] Montrul, S. Current Issues in Heritage Language Acquisition. *Annu. Rev. Appl. Linguist.* **2010**, 30, 3–23. <https://doi.org/10.1017/S0267190510000103>
- [12] Kondo-Brown, K. Heritage Language Instruction for Post-Secondary Students from Immigrant Backgrounds. *Herit. Lang. J.* **2001**, 1 (1), 1–25. <https://doi.org/10.46538/hlj.1.1.1>
- [13] Polinsky, M.; Kagan, O. Heritage Languages: In the “Wild” and in the Classroom. *Lang. Linguist. Compass* **2007**, 1 (5), 368–395. <https://doi.org/10.1111/j.1749-818X.2007.00022.x>
- [14] Valdés, G. Heritage Language Students: Profiles and Possibilities. In *Heritage Languages in America: Preserving a National Resource*; Peyton, J. K.; Ranard, D. A.; McGinnis, S., Eds.; Center for Applied Linguistics: Washington, DC, 2001; pp 37–77.
- [15] He, A. W.; Xiao, Y., Eds. *Chinese as a Heritage Language: Fostering Rooted World Citizenry*; National Foreign Language Resource Center, University of Hawai‘i: Honolulu, HI, 2008.
- [16] Knoch, U.; Chapelle, C. A. Validation of Rating Processes within an Argument-Based Framework. *Lang. Test.* **2018**, 35 (4), 477–499. <https://doi.org/10.1177/0265532217710049>
- [17] East, M. Evaluating the Reliability of a Detailed Analytic Scoring Rubric for Foreign Language Writing. *Assess. Writ.* **2009**, 14 (2), 88–115. <https://doi.org/10.1016/j.asw.2009.04.001>
- [18] Turner, C. E.; Purpura, J. E. Learning-Oriented Assessment in Second and Foreign Language Classrooms. In *Handbook of Second Language Assessment*; De Gruyter Mouton: Berlin, 2016; pp 255–274. <https://doi.org/10.1515/9781614513827-018>
- [19] Bachman, L. F.; Palmer, A. S. *Language Assessment in Practice*; Oxford University Press: Oxford, 2010.
- [20] Ke, C. A Model of Formative Task-Based Language Assessment for Chinese as a Foreign Language. *Lang. Assess. Q.* **2006**, 3 (2), 207–227. https://doi.org/10.1207/S15434311LAQ0302_6